

Tipos de datos estructurados

Ricardo Pérez López

IES Doñana, curso 2022/2023



Generado el 2022/12/15 a las 11:35:00

1. Introducción
2. Secuencias
3. Estructuras no secuenciales

1. Introducción

1.1 Composición

1.2 Conceptos básicos

1.3 Clasificación

1.4 *Hashables*

1.5 Iterables

1.6 Iteradores

1.1. Composición

Composición

- ▶ Hasta ahora, hemos aprendido que:
 - Un programa está compuesto por *instrucciones*.
 - Las instrucciones de un programa son las *expresiones* y las *sentencias*.
- ▶ Además, hemos visto que podemos crear instrucciones más complejas a partir de otras más simples. Es decir:
 - Podemos crear *expresiones más complejas* combinando entre sí expresiones más simples.
 - Podemos crear *sentencias compuestas* (estructuras de control, como bloques, condicionales, bucles, etc.) combinando entre sí otras sentencias.
- ▶ La propiedad que tienen los lenguajes de programación de crear elementos más complejos combinando otros más simples se denomina **composición**.

- ▶ La **abstracción** y la **composición** son dos conceptos relacionados:
 - *Componer* consiste en combinar elementos entre sí para formar otros más complejos.
 - *Abstraer* consiste en coger un elemento (normalmente complejo), darle un nombre y ocultar sus detalles internos (es decir, los elementos que lo componen) dentro de una caja negra.
- ▶ Lo interesante es que la combinación y la abstracción son dos mecanismos *recursivos*:
 - Podemos crear elementos complejos a partir de otros elementos complejos.
 - Podemos crear abstracciones a partir de otras abstracciones.
- ▶ Además, por supuesto, podemos crear abstracciones a partir de composiciones y composiciones a partir de abstracciones.

- ▶ Por ahora, esos conceptos (*composición* y *abstracción*) sólo los hemos aplicado a las **instrucciones** del programa:
 - La **composición de instrucciones** da lugar a las **expresiones compuestas** y a las **sentencias compuestas** (también llamadas **estructuras de control**: *secuencia, selección e iteración*).
 - La **abstracción de instrucciones** da lugar a las **abstracciones funcionales**.
- ▶ Pero también se pueden aplicar a los **datos**:
 - La **composición de datos** da lugar a los **datos compuestos** (también llamados **datos estructurados**) y, en consecuencia, a los **tipos de datos compuestos** (también llamados **tipos de datos estructurados**).
 - La **abstracción de datos** da lugar a los **datos abstractos** y, en consecuencia, a los **tipos abstractos de datos**.
- ▶ En esta unidad hablaremos de la composición de datos y dejaremos la abstracción de datos para una unidad posterior.

1.2. Conceptos básicos

Conceptos básicos

- ▶ Un **dato estructurado** (también llamado **dato compuesto**, **colección** o **contenedor**) es un dato formado, a su vez, por otros datos llamados **componentes** o **elementos**, los cuales representan su **contenido**.
- ▶ Por contra, los datos no estructurados se denominan **datos elementales**, **escalares** o **atómicos**.
- ▶ Un **tipo de dato estructurado**, también llamado **tipo compuesto**, es aquel cuyos valores son datos estructurados.
- ▶ Normalmente, se puede **acceder** de manera individual a los elementos que componen un dato estructurado y, a veces, también se pueden **modificar** esos elementos de manera individual.
- ▶ El término **estructura de datos** se suele usar como sinónimo de **tipo de dato estructurado**, aunque nosotros haremos una distinción:
 - Usaremos **tipo de dato estructurado** cuando usemos un dato sin conocer sus detalles internos de implementación.
 - Usaremos **estructura de datos** cuando nos interesen esos detalles.

1.3. Clasificación

Clasificación

- ▶ Los **datos estructurados** se pueden clasificar atendiendo a su *secuencialidad* y a su *mutabilidad*.
- ▶ Según su **secuencialidad**:

- **Secuenciales:** Son aquellos en los que se puede acceder directamente y de forma eficiente a cada uno de sus elementos indicando la posición que ocupan dentro de la secuencia.

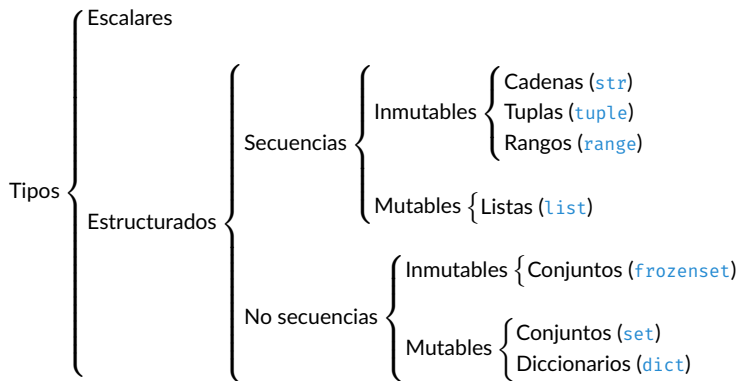
Por tanto, son colecciones *ordenadas*, ya que sus elementos están ordenados dentro de la secuencia según la posición en la que se encuentran situados dentro de la misma.

- **No secuenciales:** Son aquellos en los que **NO** se puede acceder directamente y de forma eficiente a cada uno de sus elementos indicando la posición que ocupan dentro de la colección.

En general, las estructuras no secuenciales son colecciones *desordenadas*, en las que no se puede afirmar que sus elementos se encuentran en una posición determinada dentro de la colección.

- ▶ Según su **mutabilidad**:
 - **Inmutables**: el dato estructurado no puede cambiar nunca su estado interno a lo largo de su vida.
 - **Mutables**: el dato estructurado puede cambiar su estado interno a lo largo de su vida sin cambiar su identidad.
- ▶ El **contenido** de un dato estructurado forma parte del **estado interno** de ese dato estructurado, por lo que cambiar el contenido de un dato estructurado supone cambiar también su estado interno.
- ▶ Por ejemplo, si en la lista [7, 8, 9] sustituimos su segundo elemento (el 8) por un 5 para obtener la lista [7, 5, 9], estamos cambiando el contenido de la lista y, por consiguiente, su estado interno.

Su identidad no ha cambiado, pero su estado interno sí.



1.4. Hashables

Hashables

- ▶ Un dato es *hashable* si cumple las siguientes dos condiciones:
 1. Puede compararse con otros datos usando el operador `==`.
 2. Tiene asociado un número entero llamado **hash** que nunca cambia durante toda la vida del dato.

Para obtener el *hash* de un dato, se usa la función `hash`:

- Si un dato *d* es *hashable*, `hash(d)` devolverá el *hash* de *d*.
 - En caso contrario, lanzará una excepción de tipo `TypeError`.
- ▶ Si dos datos *hashables* son iguales, entonces deben tener el mismo *hash*:

Si `x == y`, entonces debe cumplirse que `hash(x) == hash(y)`.

En cambio, si dos datos son distintos, sus *hash* no tienen por qué serlo.

► Ejemplos:

```
>>> hash('hola')
6290906884732116299
>>> hash('hola')
6290906884732116299
>>> hash(5)
5
>>> hash((1, 2, 3))
529344067295497451
>>> hash([1, 2, 3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unhashable type: 'list'
```


- ▶ El concepto de *hashable* es importante en Python ya que existen tipos de datos estructurados que sólo pueden contener elementos *hashables*.
- ▶ Por ejemplo, los elementos de un conjunto y las claves de un diccionario deben ser *hashables*.
- ▶ La mayoría de los datos inmutables predefinidos en Python son *hashables*.
- ▶ Las **colecciones inmutables** (como las tuplas o los `frozensets`) sólo son *hashables* si sus elementos también lo son.
- ▶ Las **coleccion mutables** (como las listas o los diccionarios) **NO** son *hashables*.

- El **hash** de un dato depende del estado interno del dato, ya que se calcula a partir de dicho estado interno usando un algoritmo que no nos debe preocupar por ahora.

Como el estado interno de una colección viene determinado principalmente por los elementos que contiene, el *hash* de una colección dependerá también del contenido de la colección.

Y por esta razón, las colecciones mutables no son *hashables*: si una colección es mutable, su contenido puede cambiar y, por tanto, su *hash* también cambiaría, pero esto está prohibido.

El *hash* de un dato **se calcula** en función del **estado interno** del dato y, en caso de ser una colección, también en función de su **contenido**.

- El *hash* de un dato es un número que **representa al dato y a todo su contenido**.
- En cierto modo, ese número **resume el estado del dato** en un simple número entero.

- ▶ El *hash* de un dato se utiliza internamente para acceder al dato dentro de una colección de forma directa y eficiente.
- ▶ Para ello, el intérprete utiliza ciertas técnicas que permiten localizar directamente a un dato dentro de una colección, de forma casi inmediata y sin importar el tamaño de la colección (pero recordemos que para ello es necesario que el *hash* del dato nunca cambie).
- ▶ De no usar estas técnicas, el intérprete tendría que buscar el dato secuencialmente dentro de la colección, recorriéndola desde el principio hasta el final, lo que sería mucho más lento y consumiría un tiempo que sería mayor cuanto más grande fuese la colección.

Los *hash* **permiten el acceso *directo* a un dato** dentro de una colección.

- ▶ Muy en resumen, las técnicas se basan en dividir el espacio de memoria que ocupa la colección en una serie de *contenedores* llamados **buckets**.
- ▶ Cada *bucket* va numerado por un posible valor de *hash*, de forma que el *bucket* número n contendrá todos los elementos cuyo *hash* valga n .
- ▶ Por tanto, el algoritmo que usa el intérprete para encontrar un elemento *hashable* dentro de una colección es:
 1. Calcular el *hash* del elemento a localizar.
 2. Irse directamente al *bucket* numerado con ese valor de *hash* (esta es una operación inmediata, con coste $O(1)$).
 3. Localizar dentro del *bucket* el elemento que se está buscando usando el `==`, lo cual consumirá un tiempo que, en general, no será mucho, ya que los elementos están repartidos entre todos los *buckets* y, por tanto, normalmente no habrá muchos elementos en cada *bucket*.
- ▶ Al final, se consigue encontrar al elemento (si está) de forma muy rápida, con un coste que es casi constante, independientemente de la cantidad de elementos que haya en la colección.

No se debe confundir el `id` de un dato con el `hash` de un dato:

Función <code>id</code>	Función <code>hash</code>
<ul style="list-style-type: none">▶ Devuelve la identidad de un dato.▶ Todos los datos tienen identidad.▶ Puede haber datos iguales pero no idénticos.▶ Su valor no depende del estado interno del dato y, por tanto, tampoco de su contenido.▶ Por tanto, no cambia si se modifica el dato.	<ul style="list-style-type: none">▶ Devuelve el <i>hash</i> de un dato, si es <i>hashable</i>.▶ No todos los datos son <i>hashables</i>.▶ Si dos datos son iguales, sus <i>hash</i> también deben serlo.▶ Su valor se obtiene a partir del estado interno del dato (y, por tanto, de su contenido), usando una fórmula matemática.▶ Por tanto, un dato mutable no puede ser <i>hashable</i>, ya que su <i>hash</i> cambiaría al cambiar su contenido o estado interno.

1.5. Iterables

Iterables

- ▶ Se dice que un dato compuesto es **iterable** cuando se puede acceder a todos sus elementos de uno en uno, operación que se denomina **recorrer** el iterable.
- ▶ Gracias a esto, se dice que un iterable nos permite *visitar* sus elementos o, también, *iterar* sobre sus elementos.
- ▶ Como iterables tenemos:
 - Todas las secuencias: listas, cadenas, tuplas y rangos.
 - Estructuras no secuenciales: diccionarios y conjuntos.
- ▶ Los iterables no representan un tipo concreto, sino más bien una *familia* de tipos que comparten la misma propiedad.
- ▶ Muchas funciones, como `map` y `filter`, actúan sobre iterables en general, en lugar de hacerlo sobre un tipo concreto (lista, tupla, ...).

- ▶ Por ejemplo, las listas son iterables ya que nos permite acceder a todos sus elementos de uno en uno y, por tanto, podemos recorrerla.
- ▶ Para *visitar* sus elementos podemos usar la *indexación*, y para *recorrer* toda la lista podemos usar un *bucle*:

```
def recorrer_lista(l):  
    i = 0  
    while i < len(l):  
        print(l[i])  
        i += 1
```


1.6. Iteradores

1.6.1 El bucle `for`

1.6.2 El módulo `itertools`

Iteradores

- ▶ La forma básica de recorrer un dato iterable es usando un **iterador**.
De hecho, técnicamente, un *iterable* **se define** como aquel dato al que le podemos asociar, al menos, un **iterador**.
- ▶ Un **iterador** es un objeto que sabe cómo **recorrer** un iterable.
- ▶ Para ello, el iterador crea un flujo de datos *perezoso* que va entregando los elementos del iterable de uno en uno.
- ▶ Los sucesivos elementos del flujo de datos se van obteniendo al llamar repetidamente a la función `next` aplicada al iterador.
- ▶ Cuando ya no hay más elementos disponibles, la función `next` lanza una excepción de tipo `StopIteration`, lo que indica que el iterador **se ha agotado** (se han consumido todos sus elementos), por lo que si se sigue llamando a la función `next` se seguirá lanzando esa excepción.

- ▶ Se puede obtener un iterador a partir de cualquier dato iterable aplicando la función `iter` al iterable.
(Recordemos que todo iterable debe tener asociado un iterador.)
- ▶ Ejemplo de uso de `iter` y `next`:

```
>>> lista = [1, 2, 3]
>>> it = iter(lista)
>>> next(it)
1
>>> next(it)
2
>>> next(it)
3
>>> next(it)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
StopIteration
```

- ▶ Si se le pasa un dato no iterable, `iter` lanza una excepción `TypeError`:

```
>>> it = iter(9)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'int' object is not iterable
```

- ▶ Los iteradores son **iterables perezosos de un solo uso**:
 - Son **perezosos** porque van generando sus elementos a medida que los va entregando, en lugar de generarlos todos a la vez primero.
 - Son **de un solo uso** porque cada elemento sólo se entrega una vez.
- ▶ Además, los iteradores son **iterables que actúan como sus propios iteradores**:
 - Por tanto, cuando llamamos a `iter` pasándole un iterador, se devuelve el mismo iterador:

```
>>> lista = [1, 2, 3, 4]
>>> it = iter(lista)
>>> it
<list_iterator object at 0x7f3c49aa9080>
>>> it2 = iter(it)
>>> it2
<list_iterator object at 0x7f3c49aa9080>
>>> it == it2
True
```

- En consecuencia, podemos usar un iterador en cualquier sitio donde se espere un iterable.

- Funciones como `map` y `filter` devuelven iteradores porque, al ser perezosos, son más eficientes en memoria que si devolvieran toda una lista o tupla.

Por ejemplo: ¿qué ocurre si sólo necesitamos los primeros elementos del resultado de un `map`?

- Los iteradores se pueden convertir en listas o tuplas usando las funciones `list` y `tuple`:

```
>>> l = [1, 2, 3]
>>> iterador = iter(l)
>>> t = tuple(iterador)
>>> t
(1, 2, 3)
```

- ▶ Las **expresiones generadoras**, ya conocidas por nosotros, también son expresiones que **devuelven un iterador**:

`<expr_gen> ::= (<expresión> for <identificador> in <secuencia> [if <condición>])`

- ▶ Ejemplo:

```
>>> cuadrados = (x ** 2 for x in range(1, 10))
>>> cuadrados
<generator object <genexpr> at 0x7f6a0fc7db48>
>>> next(cuadrados)
1
>>> next(cuadrados)
4
>>> next(cuadrados)
9
```

El bucle **for**

- Probablemente, la mejor forma de recorrer los elementos que devuelve un iterador es mediante una **estructura de control** llamada **bucle for**.
- Su sintaxis es:

```
for <variable>(, <variable>)* in <iterable>:  
    <sentencia>
```

que no es más que azúcar sintáctico para el siguiente código equivalente:

```
iterador = iter(<iterable>)  
while True:  
    try:  
        <variable>(, <variable>)* = next(iterador)  
    except StopIteration:  
        break  
    else:  
        <sentencia>
```

► Ejemplos:

```
for i in range(0, 4):  
    print(i)
```

devuelve:

0
1
2
3

```
for x in ['hola', 23.5, 10, [1, 2]]:  
    print(x * 2)
```

devuelve:

'holahola'
47.0
20
[1, 2, 1, 2]

- ▶ Al recorrer el iterable, la variable va almacenando en cada iteración del bucle el valor del elemento que en ese momento se está visitando.
- ▶ Si estamos recorriendo una secuencia y necesitamos recuperar tanto el valor como el **índice** de cada elemento, podemos usar la función `enumerate`.
- ▶ Esta función devuelve un iterador que va generando tuplas que contienen, además del elemento, el valor correspondiente de un contador numérico.
- ▶ Las tuplas que devuelve el iterador llevan el contador en la primera posición y el elemento de la secuencia en la segunda posición.
- ▶ Por defecto, el contador empieza desde 0 y se va incrementando de uno en uno, por lo que coincide con el índice del elemento en la secuencia:

```
>>> for i, e in enumerate(['a', 'b', 'c']):  
...     print('El elemento en la posición ' + str(i) + ' es ' + str(e))  
...  
El elemento en la posición 0 es a  
El elemento en la posición 1 es b  
El elemento en la posición 2 es c
```

- ▶ Existen iterables e iteradores incluso donde uno menos se lo podría esperar.
- ▶ Por ejemplo, **los archivos abiertos también son iterables**, ya que se pueden recorrer línea a línea usando un iterador:

```
with open('archivo.txt') as f:  
    for linea in f:  
        print(linea)
```

- ▶ Esta forma de recorrer los archivos, además de resultar simple y elegante, también resulta muy eficiente, ya que se va recuperando cada línea de una en una en lugar de todas a la vez.

El módulo `itertools`

- ▶ El módulo `itertools` contiene una variedad de iteradores de uso frecuente, así como funciones que combinan varios iteradores.
- ▶ Algunos de esos iteradores son muy especiales porque pueden devolver flujos infinitos o valores que se repiten continuamente, lo cual contradice en cierta manera lo que dijimos cuando definimos los iteradores como «*iterables de un solo uso*».
- ▶ `itertools.count([<inicio>[, <paso>]])` devuelve un flujo infinito de valores separados uniformemente. Se puede indicar opcionalmente un valor de comienzo (que por defecto es `0`) y el intervalo entre números (que por defecto es `1`):

`itertools.count()` \Rightarrow 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ...

`itertools.count(10)` \Rightarrow 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ...

`itertools.count(10, 5)` \Rightarrow 10, 15, 20, 25, 30, 35, 40, 45, ...

- ▶ `itertools.cycle(<iterador>)` devuelve un nuevo iterador que va generando sus elementos del primero al último, repitiéndolos indefinidamente:

`itertools.cycle([1, 2, 3, 4])` \Rightarrow 1, 2, 3, 4, 1, 2, 3, 4, ...

- ▶ `itertools.repeat(<elem>[, <n>])` devuelve `<n>` veces el elemento `<elem>`, o lo devuelve indefinidamente si no se indica `<n>`:

`itertools.repeat('abc')` \Rightarrow abc, abc, abc, abc, abc, abc, abc, ...

`itertools.repeat('abc', 5)` \Rightarrow abc, abc, abc, abc, abc

2. Secuencias

2.1 Concepto de secuencia

2.2 Operaciones comunes

2.3 Inmutables

2.4 Mutables

2.1. Concepto de secuencia

Concepto de secuencia

- ▶ Una **secuencia** s es un dato estructurado *iterable* que cumple lo siguiente:
 1. Se le puede calcular su longitud (la cantidad de elementos que contiene) mediante la función `len`.
 2. Cada elemento que contiene lleva asociado un número entero llamado **índice**, comprendido entre `0` y `len(s) - 1`.
 3. Permite el acceso eficiente a cada uno de sus elementos mediante indexación `s[i]`, siendo i el índice del elemento.
- ▶ Las secuencias se dividen en:
 - **Inmutables:** cadenas (`str`), tuplas (`tuple`) y rangos (`range`).
 - **Mutables:** listas (`list`)

2.2. Operaciones comunes

Operaciones comunes

- ▶ Todas las secuencias (ya sean cadenas, listas, tuplas o rangos) comparten un conjunto de **operaciones comunes**.
- ▶ Los rangos son una excepción, ya que sus elementos se crean a partir de una fórmula y, por eso, no admiten ni la concatenación ni la repetición.
- ▶ La siguiente tabla recoge las operaciones comunes sobre secuencias, ordenadas por prioridad ascendente. \underline{s} y \underline{t} son secuencias del mismo tipo, \underline{n} , \underline{i} , \underline{j} y \underline{k} son enteros y \underline{x} es un dato cualquiera que cumple con las restricciones que impone \underline{s} .

Operación	Resultado
<code>x in s</code>	<code>True</code> si algún elemento de <u>s</u> es igual a <u>x</u>
<code>x not in s</code>	<code>False</code> si algún elemento de <u>s</u> es igual a <u>x</u>
<code>s + t</code>	La concatenación de <u>s</u> y <u>t</u> (no va con rangos)
<code>s * n</code> <code>n * s</code>	(Repetición) Equivale a concatenar <u>s</u> consigo misma <u>n</u> veces (no va con rangos)
<code>s[i]</code>	El <u>i</u> -ésimo elemento de <u>s</u> , empezando por 0
<code>s[i:j]</code>	Rodaja de <u>s</u> desde <u>i</u> hasta <u>j</u>
<code>s[i:j:k]</code>	Rodaja de <u>s</u> desde <u>i</u> hasta <u>j</u> con paso <u>k</u>
<code>len(s)</code>	Longitud de <u>s</u>
<code>min(s)</code>	El elemento más pequeño de <u>s</u>
<code>max(s)</code>	El elemento más grande de <u>s</u>
<code>s.index(x[, i[, j]])</code>	El índice de la primera aparición de <u>x</u> en <u>s</u> (desde el índice <u>i</u> inclusive y antes del <u>j</u>)
<code>s.count(x)</code>	Número total de apariciones de <u>x</u> en <u>s</u>

- ▶ Además de estas operaciones, las secuencias admiten **comparaciones** con los operadores `==`, `!=`, `<`, `<=`, `>` y `>=`.
- ▶ Dos secuencias `s` y `t` son iguales (`s == t`) si:
 - Son del mismo tipo (`type(s) == type(t)`).
 - Tienen la misma longitud (`len(s) == len(t)`).
 - Contienen los mismos elementos en el mismo orden (`s[0] == t[0]`, `s[1] == t[1]`, etcétera).
- ▶ Por supuesto, las dos secuencias son distintas (`s != t`) si no son iguales.
- ▶ Se pueden comparar dos secuencias con los operadores `<`, `<=`, `>` y `>=` para comprobar si una es menor (o igual) o mayor (o igual) que la otra si:
 - Son del mismo tipo (si no son del mismo tipo, lanza una excepción).
 - No son rangos.

- ▶ Las comparaciones $<$, $<=$, $>$ y $>=$ se hacen lexicográficamente elemento a elemento, como en un diccionario.
- ▶ Por ejemplo, `'adios' < 'hola'` porque `adios` aparece antes que `hola` en el diccionario.
- ▶ Con el resto de las secuencias se actúa igual que con las cadenas.

Dadas dos secuencias \underline{s} y \underline{t} , para ver si $\underline{s} < \underline{t}$ se procede así:

- Se empieza comparando el primer elemento de \underline{s} con el primero de \underline{t} .
 - Si son iguales, se pasa al siguiente hasta encontrar algún elemento de \underline{s} que sea distinto a su correspondiente de \underline{t} .
 - Si llegamos al final de \underline{s} sin haber encontrado ningún elemento distinto a su correspondiente en \underline{t} , es porque $\underline{s} == \underline{t}$.
 - En cuanto se encuentre un elemento de \underline{s} que no es igual a su correspondiente de \underline{t} , se comparan esos elementos y se devuelve el resultado de esa comparación.
- ▶ Los rangos no se pueden comparar con $<$, $<=$, $>$ o $>=$.

► Ejemplos:

```
>>> (1, 2, 3) == (1, 2, 3)
True
>>> (1, 2, 3) != (1, 2, 3)
False
>>> (1, 2, 3) == (3, 2, 1)
False
>>> (1, 2, 3) < (3, 2, 1)
True
>>> (1, 2, 3) < (1, 2, 4)
True
>>> (1, 2, 3) < (1, 2, 4, 5)
True
>>> 'hola' < 'adios'
False
>>> range(0, 3) < range(3, 6)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: '<' not supported between instances of 'range' and 'range'
```

2.3. Inmutables

2.3.1 Cadenas (`str`)

2.3.2 Tuplas

2.3.3 Rangos

Cadenas (str)

- ▶ Las **cadenas** son secuencias inmutables y *hashables* de caracteres.
- ▶ No olvidemos que en Python no existe el tipo *carácter*. En Python, un carácter es una cadena de longitud 1.
- ▶ Las cadenas literales se pueden crear:
 - Con comillas simples (') o dobles ("):

```
>>> 'hola'
'hola'
>>> "hola"
'hola'
```

- Con triples comillas (''' o """):

```
>>> """hola
... qué tal"""
'hola\nqué tal'
```

- ▶ Las cadenas implementan todas las operaciones de las secuencias, además de los métodos que se pueden consultar en <https://docs.python.org/3/library/stdtypes.html#string-methods>

Formateado de cadenas

- ▶ Una **cadena formateada** (también llamada ***f-string***) es una cadena literal que lleva un prefijo **f** o **F**.
- ▶ Estas cadenas contienen **campos de sustitución**, que son expresiones encerradas entre llaves.
- ▶ En realidad, las cadenas formateadas son expresiones evaluadas en tiempo de ejecución.

► Sintaxis:

```
⟨f_string⟩ ::= (⟨carácter_literal⟩ | { { | } } | ⟨sustitución⟩)*  
⟨sustitución⟩ ::= { ⟨expresión⟩ [!⟨conversión⟩] [:⟨especif⟩] }  
⟨conversión⟩ ::= s | r | a  
⟨especif⟩ ::= (⟨carácter_literal⟩ | NULL | ⟨sustitución⟩)*  
⟨carácter_literal⟩ ::= ⟨cualquier carácter Unicode excepto {, } o NULL⟩
```

- Las partes de la cadena que van fuera de las llaves se tratan literalmente, excepto las dobles llaves { { y } }, que son sustituidas por una sola llave.
- Una { marca el comienzo de un **campo de sustitución** (⟨sustitución⟩), que empieza con una expresión.
- Tras la expresión puede venir un **conversión** (⟨conversión⟩), introducida por una exclamación !.
- También puede añadirse un **especificador de formato** (⟨especif⟩) después de dos puntos :.
- El campo de sustitución termina con una }.

- ▶ Las expresiones en un literal de cadena formateada son tratadas como cualquier otra expresión Python encerrada entre paréntesis, con algunas excepciones:
 - No se permiten expresiones vacías.
 - Las expresiones lambda deben ir entre paréntesis.
- ▶ Los campos de sustitución pueden contener saltos de línea pero no comentarios.
- ▶ Si se indica una conversión, el resultado de evaluar la expresión se convierte antes de aplicar el formateado.

La conversión `!s` llama a la función `str` sobre el resultado, `!r` llama a `repr` y `!a` llama a `ascii`.

- ▶ A continuación, el resultado es formateado usando la función `format`.
- ▶ Finalmente, el resultado del formateado es incluido en el valor final de la cadena completa.

- La sintaxis general de un especificador de formato es:

```

<especif> ::= [[<relleno>]<alig>][<signo>][#][0][<ancho>][<grupos>][.<precision>][<tipo>]
<relleno> ::= <cualquier carácter>
<alig> ::= < > | = | ^
<signo> ::= + | - | <espacio>
<ancho> ::= <dígito>+
<grupos> ::= _ | ,
<precision> ::= <dígito>+
<tipo> ::= b | c | d | e | E | f | F | g | G | n | o | s | x | X | %

```

- Los especificadores de formato de nivel superior pueden incluir campos de sustitución anidados.
- Estos campos anidados pueden incluir, a su vez, sus propios campos de conversión y sus propios especificadores de formato, pero no pueden incluir más campos de sustitución anidados.
- Para más información, consultar <https://docs.python.org/3.7/library/string.html#format-specification-mini-language>

► Ejemplos de cadenas formateadas:

```
>>> nombre = 'Pepe'
>>> f'El nombre es: {nombre}'           # Se sustituye la variable por su valor
'El nombre es: Pepe'
>>> apellidos = 'Pérez'
>>> f'El nombre es: {nombre} {apellidos}' # Igual
'El nombre es: Pepe Pérez'
>>> f'El nombre es: {nombre + apellidos}' # Se puede usar cualquier expresión
'El nombre es: Pepe Pérez'
>>> f'Formato con anchura: {nombre:10}'  # Las cadenas se alinean a la izquierda
'Formato con anchura: Pepe          '
>>> f'Formato con anchura: {nombre:<10}'  # Igual que lo anterior
'Formato con anchura: Pepe          '
>>> f'Formato con anchura: {nombre:>10}'  # Alinea a la derecha
'Formato con anchura:      Pepe'
>>> f'Formato con anchura: {nombre:^10}'  # Alinea al centro
'Formato con anchura:   Pepe   '
```

► Ejemplos de cadenas formateadas con números positivos:

```
>>> x, y = 400, 300
>>> f'La suma de {x} y {y} es {x + y}' # Se puede usar cualquier expresión
'La suma de 400 y 300 es 700'
>>> f'Formato con anchura: {x:10}'      # Los números se alinean a la derecha
'Formato con anchura:      400'
>>> f'Formato con anchura: {x:>10}'      # Igual que lo anterior
'Formato con anchura:      400'
>>> f'Formato con anchura: {x:2}'        # Ancho demasiado pequeño, se ignora
'Formato con anchura: 400'
>>> f'Formato con anchura: {x:<10}'      # Alinea a la izquierda
'Formato con anchura: 400      '
>>> f'Formato con anchura: {x:>10}'      # Alinea a la derecha
'Formato con anchura:      400'
>>> f'Formato con anchura: {x:^10}'      # Alinea al centro
'Formato con anchura:   400    '
>>> f'Formato con anchura: {x:=10}'      # En positivos no hay diferencia con >
'Formato con anchura:      400'
>>> f'Formato con anchura: {x:@<10}'    # Alinea a la izquierda, rellena con @
'Formato con anchura: 400@@@@@@@@'
>>> f'Formato con anchura: {x:@>10}'    # Alinea a la derecha, rellena con @
'Formato con anchura: @@@@@@@@@400'
>>> f'Formato con anchura: {x:@^10}'    # Alinea al centro, rellena con @
'Formato con anchura: @@@@400@@@@'
>>> f'Formato con anchura: {x:@=10}'    # En positivos no hay diferencia con >
'Formato con anchura: @@@@@@@@@400'
```

► Ejemplos de cadenas formateadas con números negativos:

```
>>> z = -400
>>> f'Formato con anchura: {z:10}' # A la derecha, signo junto al nº
'Formato con anchura:      -400'
>>> f'Formato con anchura: {z:<10}' # A la izquierda, signo junto al nº
'Formato con anchura: -400      '
>>> f'Formato con anchura: {z:>10}' # A la derecha, signo junto al nº
'Formato con anchura:      -400'
>>> f'Formato con anchura: {z:^10}' # Al centro, signo junto al nº
'Formato con anchura:   -400   '
>>> f'Formato con anchura: {z:=10}' # A la derecha, signo junto al relleno
'Formato con anchura: -      400'
>>> f'Formato con anchura: {z:010}' # A la derecha, signo junto al relleno
'Formato con anchura: -000000400'
>>> f'Formato con anchura: {z:0<10}' # A la izquierda, signo junto al nº
'Formato con anchura: -400000000'
>>> f'Formato con anchura: {z:0>10}' # A la derecha, signo junto al nº
'Formato con anchura: 000000-400'
>>> f'Formato con anchura: {z:0^10}' # Al centro, signo junto al nº
'Formato con anchura: 000-400000'
>>> f'Formato con anchura: {z:0=10}' # A la derecha, signo junto al relleno
'Formato con anchura: -000000400'
>>> f'Formato con anchura: {z:ð<10}' # A la izquierda, signo junto al nº
'Formato con anchura: -400ðððððð'
>>> f'Formato con anchura: {z:ð>10}' # A la derecha, signo junto al nº
'Formato con anchura: ðððððð-400'
>>> f'Formato con anchura: {z:ð^10}' # Al centro, signo junto al nº
'Formato con anchura: ððð-400ððð'
>>> f'Formato con anchura: {z:ð=10}' # A la derecha, signo junto al relleno
'Formato con anchura: -ðððððð400'
```

► Ejemplos de cadenas formateadas con números en coma flotante:

```
>>> from math import pi
>>> f'El valor de pi es {pi:6.3}'      # Ancho 6, precisión 3
'El valor de pi es    3.14'
>>> f'El valor de pi es {pi:10.3}'    # Ancho 10, precisión 3
'El valor de pi es    3.14'
>>> f'El valor de pi es {pi:<10.3}'   # A la izquierda
'El valor de pi es 3.14   '
>>> f'El valor de pi es {pi:>10.3}'    # A la derecha
'El valor de pi es    3.14'
>>> f'El valor de pi es {pi:^10.3}'    # Al centro
'El valor de pi es   3.14   '
>>> f'El valor de pi es {pi:=10.3}'    # A la derecha
'El valor de pi es    3.14'
>>> f'El valor de pi es {pi:10.3f}'    # 3 dígitos en la parte fraccionaria
'El valor de pi es    3.142'
>>> f'El valor de pi es {pi:<10.3f}'   # A la izquierda
'El valor de pi es 3.142   '
>>> f'El valor de pi es {pi:>10.3f}'    # A la derecha
'El valor de pi es    3.142'
>>> f'El valor de pi es {pi:^10.3f}'    # Al centro
'El valor de pi es   3.142   '
>>> f'El valor de pi es {pi:=10.3f}'    # A la derecha
'El valor de pi es    3.142'
>>> f'El valor de pi es {-pi:=10.3f}'  # Los negativos, igual que los enteros
'El valor de pi es -    3.142'
```

► Más ejemplos:

```
>>> nombre = "Fred"
>>> f"Dice que su nombre es {nombre!r}."
"Dice que su nombre es 'Fred'."
>>> f"Dice que su nombre es {repr(nombre)}." # repr es equivalente a !r
"Dice que su nombre es 'Fred'."
>>> ancho = 10
>>> precision = 4
>>> value = decimal.Decimal("12.34567")
>>> f"result: {value:{ancho}.{precision}}" # campos anidados
'result:      12.35'
>>> import datetime
>>> hoy = datetime.datetime(year=2017, month=1, day=27)
>>> f"{hoy:%B %d, %Y}" # usando especificador de formato de fecha
'January 27, 2017'
>>> numero = 1024
>>> f"{numero:#0x}" # usando especificador de formato de enteros
'0x400'
```


Expresiones regulares

- ▶ Las **expresiones regulares** (también llamados *regex*) constituyen un pequeño lenguaje muy especializado incrustado dentro de Python y disponible a través del módulo [re](#).
- ▶ Usando este pequeño lenguaje es posible especificar **reglas sintácticas** de una forma distinta pero parecida a las *gramáticas EBNF* (aunque con menos poder expresivo).
- ▶ Esas reglas sintácticas se pueden usar luego para **comprobar si una cadena se ajusta a un patrón**.
- ▶ Este patrón puede ser frases en español, o direcciones de correo electrónico o cualquier otra cosa.
- ▶ A continuación, se pueden hacer preguntas del tipo: «¿Esta cadena se ajusta al patrón?» o «¿Hay algo que se ajuste al patrón en alguna parte de esta cadena?».
- ▶ También se pueden usar las *regexes* para **modificar** una cadena o **dividirla en partes** según el patrón indicado.

- ▶ El lenguaje de las expresiones regulares es relativamente pequeño y restringido, por lo que no es posible usarlo para realizar cualquier tipo de procesamiento de cadenas.
- ▶ Además, hay procesamientos que se pueden realizar con *regexes* pero las expresiones que resultan se vuelven muy complicadas.
- ▶ En estos casos, es mejor escribir directamente código Python ya que, aunque el código resultante pueda resultar más lento, probablemente resulte más fácil de leer.

- ▶ Para más información sobre cómo crear y usar expresiones regulares, consultar:
 - Tutorial de introducción en <https://docs.python.org/3/howto/regex.html>
 - Documentación del módulo `re` en <https://docs.python.org/3/library/re.html>

Tuplas

- ▶ Las **tuplas** (`tuple`) son secuencias inmutables, usadas frecuentemente para representar colecciones de datos heterogéneos (es decir, de tipos distintos).
- ▶ También se usan en aquellos casos en los que se necesita una secuencia inmutable de datos homogéneos (por ejemplo, para almacenar datos en un conjunto o un diccionario).
- ▶ Las tuplas se pueden crear así:

- Si es una tupla vacía, con paréntesis vacíos: `()`

- Si sólo tiene un elemento, se pone una coma detrás:

`a,` `(a,)`

- Si tiene más de un elemento, se separan con comas:

`a, b, c` `(a, b, c)`

- Usando la función `tuple(<iterable>)`.

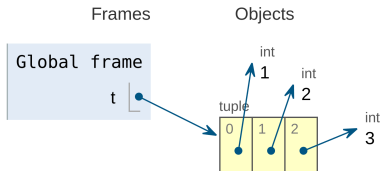
- ▶ Observar que lo que construye la tupla es realmente la coma, no los paréntesis.
- ▶ Los paréntesis son opcionales, excepto en dos casos:
 - La tupla vacía: `()`
 - Cuando son necesarios para evitar ambigüedad.

Por ejemplo, `f(a, b, c)` es una llamada a una función con tres argumentos, mientras que `f((a, b, c))` es una llamada a una función con un único argumento que es una tupla de tres elementos.
- ▶ Las tuplas implementan todas las operaciones comunes de las secuencias.
- ▶ En general, las tuplas se pueden considerar como la versión inmutable de las listas.
- ▶ Además, las tuplas son *hashables* si sus elementos también lo son.

- ▶ En memoria, las tuplas se almacenan mediante una estructura de datos donde sus elementos se identifican mediante un índice, que es un número entero que indica la posición que ocupa el elemento dentro de la tupla.
- ▶ Por ejemplo, la siguiente tupla:

```
t = (1, 2, 3)
```

se almacenaría de la siguiente forma según lo representa la herramienta Pythontutor:



Tupla almacenada en memoria

Rangos

- ▶ Los **rangos** (`range`) representan secuencias inmutables y *hashables* de números enteros y se usan frecuentemente para hacer bucles que se repitan un determinado número de veces.
- ▶ Los rangos se crean con la función `range`:

```
range([start: int,] stop: int [, step: int]) -> range
```

- ▶ Cuando se omite *start*, se entiende que es `0`.
- ▶ Cuando se omite *step*, se entiende que es `1`.
- ▶ El valor de *stop* no se alcanza nunca.
- ▶ Cuando *start* y *stop* son iguales, representa el *rango vacío*.
- ▶ *step* debe ser siempre distinto de cero.
- ▶ Cuando *start* es mayor que *stop*, el valor de *step* debería ser negativo. En caso contrario, también representaría el rango vacío.

- ▶ El **contenido** de un rango r vendrá determinado por la fórmula:

$$r[i] = start + step \cdot i$$

donde $i \geq 0$. Además:

- Si $step > 0$, se impone también la restricción $r[i] < stop$.
- Si $step < 0$, se impone también la restricción $r[i] > stop$.
- ▶ Un rango es **vacío** cuando $r[0]$ no satisface las restricciones anteriores.
- ▶ Los rangos admiten **índices negativos**, pero se interpretan como si se indexara desde el final de la secuencia usando índices positivos.
- ▶ Los rangos implementan **todas las operaciones de las secuencias, excepto la concatenación y la repetición**.

Esto es debido a que los rangos sólo pueden representar secuencias que siguen un patrón muy estricto, y las repeticiones y las concatenaciones a menudo violan ese patrón.

- ▶ **Los rangos son perezosos** y además ocupan mucha menos memoria que las listas o las tuplas (sólo hay que almacenar $start$, $stop$ y $step$).

- La **forma normal** de un rango es una expresión en la que se llama a la función `range` con los argumentos necesarios para construir el rango:

```
>>> range(10)
range(0, 10)
>>> range(0, 10)
range(0, 10)
>>> range(0, 10, 1)
range(0, 10)
```

```
>>> range(0, 30, 5)
range(0, 30, 5)
>>> range(0, -10, -1)
range(0, -10, -1)
>>> range(4, -5, -2)
range(4, -5, -2)
```

- Para ver con claridad todos los elementos de un rango, podemos convertirlo en una tupla o una lista. Por ejemplo:

```
>>> list(range(10))
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
>>> list(range(1, 11))
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
>>> list(range(0, 30, 5))
[0, 5, 10, 15, 20, 25]
>>> list(range(0, 10, 3))
[0, 3, 6, 9]
>>> list(range(0, -10, -1))
[0, -1, -2, -3, -4, -5, -6, -7, -8, -9]
>>> list(range(0))
[]
>>> list(range(1, 0))
[]
```

- ▶ Dos rangos son considerados **iguales** si representan la misma secuencia de valores, sin importar si tienen distintos valores de *start*, *stop* o *step*.
- ▶ Por ejemplo:

```
>>> range(20) == range(0, 20)
True
>>> range(0, 20) == range(0, 20, 2)
False
>>> range(0, 3, 2) == range(0, 4, 2)
True
>>> range(0) == range(2, 1, 3)
True
```

- El **rango vacío** es un valor que no tiene expresión canónica, ya que cualquiera de las siguientes expresiones representan al rango vacío tan bien como cualquier otra:
 - `range(0)`.
 - `range(a, a)`, donde a es cualquier entero.
 - `range(a, b, c)`, donde $a \geq b$ y $c > 0$.
 - `range(a, b, c)`, donde $a \leq b$ y $c < 0$.

```
>>> range(3, 3) == range(4, 4)
True
>>> range(4, 3) == range(3, 4, -1)
True
```

2.4. Mutables

2.4.1 Listas

2.4.2 Operaciones mutadoras

Listas

- ▶ Las **listas** son secuencias *mutables*, usadas frecuentemente para representar colecciones de elementos heterogéneos.
- ▶ Al ser mutables, las listas **no** son *hashables*.
- ▶ Se pueden construir de varias maneras:
 - Usando corchetes vacíos para representar la lista vacía: `[]`.
 - Usando corchetes y separando los elementos con comas:
`[a]`
`[a, b, c]`
 - Usando la función `list` con las sintaxis `list()` o `list(<iterable>)`.

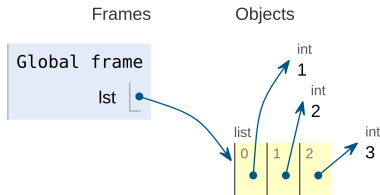
- ▶ La función `list` construye una lista cuyos elementos son los mismos (y están en el mismo orden) que los elementos de `<iterable>`.
- ▶ `<iterable>` puede ser:
 - una secuencia,
 - un contenedor sobre el que se pueda iterar, o
 - un iterador.
- ▶ Si se llama sin argumentos, devuelve una lista vacía.
- ▶ Por ejemplo:

```
>>> list('hola')  
['h', 'o', 'l', 'a']  
>>> list((1, 2, 3))  
[1, 2, 3]
```

- ▶ En memoria, las listas se almacenan mediante una estructura de datos donde sus elementos se identifican mediante un índice, que es un número entero que indica la posición que ocupa el elemento dentro de la lista.
- ▶ Por ejemplo, la siguiente lista:

```
lst = [1, 2, 3]
```

se almacenaría de la siguiente forma según lo representa la herramienta Pythontutor:



Lista almacenada en memoria

Listas por comprensión

- ▶ También se pueden crear **listas por comprensión** usando la misma sintaxis de las **expresiones generadoras** pero encerrando la expresión entre corchetes en lugar de entre paréntesis.
- ▶ Su sintaxis es:

```
<lista_comp> ::= [ <expresión> (for <identificador> in <secuencia> [if <condición>])+ ]
```

- ▶ Por ejemplo:

```
>>> [x ** 2 for x in [1, 2, 3]]  
[1, 4, 9]
```

- ▶ Como se ve, el resultado es directamente una lista, no un iterador.
- ▶ Por tanto, a diferencia de lo que pasa con las expresiones generadoras, **el resultado de una lista por comprensión no es perezoso**, cosa que habrá que tener en cuenta para evitar consumir más memoria de la necesaria o generar elementos que al final no sean necesarios.

- Por ejemplo, la siguiente expresión generadora:

```
res_gen = (x ** 2 for x in range(0, 10000000000000))
```

es mucho más eficiente en tiempo y espacio que la lista por comprensión:

```
res_list = [x ** 2 for x in range(0, 10000000000000)]
```

ya que la expresión generadora devuelve un **iterador** que irá generando los valores de uno en uno a medida que los vayamos recorriendo con `next(res_gen)`.

En cambio, la lista por comprensión genera todos los valores de la lista a la vez y los almacena todos juntos en la memoria.

- A cambio, la ventaja de tener una lista frente a tener un iterador es que podemos acceder directamente a cualquier elemento de la lista mediante la indexación.

- ▶ Las listas por comprensión, al igual que las expresiones generadoras, **determinan su propio ámbito**.
- ▶ Ese ámbito abarca toda la lista por comprensión, de principio a fin.
- ▶ Los identificadores que aparecen en la cláusula **for** se consideran *variables ligadas* en la lista por comprensión.
- ▶ Esos identificadores se van ligando, uno a uno, a cada elemento de la secuencia indicada en la cláusula **in**.

► Como son variables ligadas, cumplen estas dos propiedades:

1. Se pueden renombrar (siempre de forma consistente) sin que la lista por comprensión cambie su significado.

Por ejemplo, las dos listas por comprensión siguientes son equivalentes, puesto que producen el mismo resultado:

```
[x for x in (1, 2, 3)]
```

```
[y for y in (1, 2, 3)]
```

2. No se pueden usar fuera de la lista por comprensión, ya que estarían fuera de su ámbito y no serían visibles.

Por ejemplo, lo siguiente daría un error de nombre:

```
>>> e = [x for x in (1, 2, 3)]
>>> x      # Intento acceder a la 'x' de la lista por comprensión
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'x' is not defined
```

Operaciones mutadoras

- En la siguiente tabla, \underline{s} es una instancia de un tipo de secuencia mutable (por ejemplo, una lista), \underline{t} es cualquier dato iterable y \underline{x} es un dato cualquiera que cumple con las restricciones que impone \underline{s} :

Operación	Resultado
$s[i] = x$	El elemento i -ésimo de \underline{s} se sustituye por \underline{x}
$s[i:j] = t$	La rodaja de \underline{s} desde i hasta j se sustituye por \underline{t}
$s[i:j:k] = t$	Los elementos de $s[i:j:k]$ se sustituyen por \underline{t}
<code>del</code> $s[i]$	Elimina el elemento i -ésimo de \underline{s}
<code>del</code> $s[i:j]$	Elimina los elementos de $s[i:j]$ Equivale a hacer $s[i:j] = []$
<code>del</code> $s[i:j:k]$	Elimina los elementos de $s[i:j:k]$

Operación	Resultado
<code>s.append(x)</code>	Añade <u>x</u> al final de <u>s</u> ; es igual que <code>s[len(s):len(s)] = [x]</code>
<code>s.clear()</code>	Elimina todos los elementos de <u>s</u> ; es igual que <code>del s[:]</code>
<code>s.copy()</code>	Crea una copia <i>superficial</i> de <u>s</u> ; es igual que <code>s[:]</code>
<code>s.extend(t)</code> <code>s += t</code>	Extiende <u>s</u> con el contenido de <u>t</u> ; es igual que <code>s[len(s):len(s)] = t</code>
<code>s *= n</code>	Modifica <u>s</u> repitiendo su contenido <u>n</u> veces
<code>s.insert(i, x)</code>	Inserta <u>x</u> en <u>s</u> en el índice <u>i</u> ; es igual que <code>s[i:i] = [x]</code>
<code>s.pop([i])</code>	Extrae el elemento <u>i</u> de <u>s</u> y lo devuelve (por defecto, <u>i</u> vale <code>-1</code>)
<code>s.remove(x)</code>	Quita el primer elemento de <u>s</u> que sea igual a <u>x</u>
<code>s.reverse()</code>	Invierte los elementos de <u>s</u>
<code>s.sort()</code>	Ordena los elementos de <u>s</u>

- ▶ La **copia superficial** (a diferencia de la **copia profunda**) significa que sólo se copia el objeto sobre el que se aplica la copia, **no sus elementos**.
- ▶ Por tanto, al crear la copia superficial, se crea sólo un nuevo objeto, donde se copiarán las referencias de los elementos del objeto original.
- ▶ Esto influye, sobre todo, cuando los elementos de una colección mutable también son objetos mutables.
- ▶ Por ejemplo, si tenemos listas dentro de otra lista, y copiamos ésta última con `.copy()`, la nueva lista compartirá elementos con la lista original:

```
>>> x = [[1, 2], [3, 4]] # los elementos de «x» también son listas
>>> y = x.copy()        # «y» es una copia de «x»
>>> x is y               # no son la misma lista
False
>>> x[0] is y[0]         # sus elementos no se han copiado,
True                    # sino que están compartidos por «x» e «y»
>>> x[0].append(9)
>>> x
[[1, 2, 99], [3, 4]]
>>> y
[[1, 2, 99], [3, 4]]
```

- El método `sort` permite ordenar los elementos de la secuencia de forma ascendente o descendente:

```
>>> x = [3, 6, 2, 9, 1, 4]
>>> x.sort()
>>> x
[1, 2, 3, 4, 6, 9]
>>> x.sort(reverse=True)
>>> x
[9, 6, 4, 3, 2, 1]
```

3. Estructuras no secuenciales

3.1 Conjuntos (`set` y `frozenset`)

3.2 Dicionarios (`dict`)

3.3 Documentos XML

3.1. Conjuntos (set y frozenset)

3.1.1 Conjuntos por comprensión

3.1.2 Operaciones

3.1.3 Operaciones sobre conjuntos mutables

3.1.4 Recorrido de conjuntos

Conjuntos (set y frozenset)

- ▶ Un conjunto es una colección **no ordenada** de elementos *hashables*.
- ▶ Se usan frecuentemente para comprobar si un elemento pertenece a un grupo, para eliminar duplicados en una secuencia y para realizar operaciones matemáticas como la *unión*, la *intersección* y la *diferencia simétrica*.
- ▶ Como cualquier otra colección, los conjuntos permiten el uso de:
 - `x in c`
 - `len(c)`
 - `for x in c`
- ▶ Como son colecciones no ordenadas, los conjuntos **no almacenan la posición** de los elementos o el **orden** en el que se insertaron.
- ▶ Por tanto, tampoco admiten la indexación, las rodajas ni cualquier otro comportamiento propio de las secuencias.

- ▶ Cuando decimos que **un conjunto no está ordenado**, queremos decir que los elementos que contiene no se encuentran situados en una posición concreta.

Es lo contrario de lo que ocurre con las secuencias, donde cada elemento se encuentra en una posición indicada por su *índice* y podemos acceder a él usando la indexación.

- ▶ Además, en un conjunto **no puede haber elementos repetidos** (un elemento concreto sólo puede estar *una* vez dentro de un conjunto, es decir, o está una vez o no está).
- ▶ En resumen:

En un conjunto:

Un elemento concreto, o está una vez, o no está.

Si está, no podemos saber en qué posición (no tiene sentido preguntárselo).

- ▶ Existen dos tipos predefinidos de conjuntos: `set` y `frozenset`.
- ▶ El tipo `set` es **mutable**, es decir, que su contenido puede cambiar usando métodos como `add` y `remove`.

Como es mutable, **no es hashable** y, por tanto, no puede usarse como clave de un diccionario o como elemento de otro conjunto.

- ▶ El tipo `frozenset` es **immutable** y *hashable*. Por tanto, su contenido no se puede cambiar una vez creado y puede usarse como clave de un diccionario o como elemento de otro conjunto.

- ▶ Los dos tipos de conjuntos se crean con las funciones `set([<iterable>])` y `frozenset([<iterable>])`:
 - Si se llaman *sin argumentos*, devuelven un conjunto **vacío**:
 - `set()` devuelve un conjunto vacío de tipo `set`.
 - `frozenset()` devuelve un conjunto vacío de tipo `frozenset`.

```
>>> set()
set()
>>> frozenset()
frozenset()
```

Como se ve, esas son, precisamente, las **formas normales** de un conjunto vacío de tipo `set` y `frozenset`.

- Si se les pasa un *iterable* (como por ejemplo, una lista), devuelve un conjunto formado por los elementos del iterable:

```
>>> set([4, 3, 2, 2, 4])
{2, 3, 4}
>>> frozenset([4, 3, 2, 2, 4])
frozenset({2, 3, 4})
```

- ▶ Además, existe una *sintaxis especial* para escribir **literales de conjuntos no vacíos de tipo set**, que consiste en encerrar sus elementos entre llaves y separados por comas: `{'pepe', 'juan'}`.

```
>>> x = {'pepe', 'juan'} # un literal de tipo set, como set(['pepe', 'juan'])
>>> x
{'pepe', 'juan'}
>>> type(x)
<class 'set'>
```

Esa es, precisamente, la **forma normal** de un conjunto no vacío y, por tanto, la que se usa cuando se visualiza desde el intérprete o se imprime con `print`.

- ▶ Por tanto, para crear conjuntos congelados usando `frozenset` podemos usar esa sintaxis en lugar de usar listas como hicimos antes:

```
>>> frozenset({4, 3, 2, 2, 4})
frozenset({2, 3, 4})
```

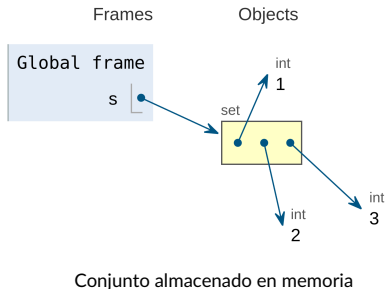
- También podría usarse con la función `set`, pero entonces estaríamos creando un nuevo conjunto igual que el anterior, aunque no idéntico (es decir, sería una *copia* del original):

```
>>> s = {4, 3, 2, 2, 4}
>>> s
{2, 3, 4}
>>> t = set(s)
>>> t
{2, 3, 4}
>>> s == t
True
>>> s is t
False
```

- ▶ En memoria, los conjuntos se almacenan mediante una estructura de datos donde sus elementos no se identifican mediante ningún índice o clave especial.
- ▶ Por ejemplo, el siguiente conjunto:

```
d = {1, 2, 3}
```

se almacenaría de la siguiente forma según lo representa la herramienta Pythontutor:



Conjuntos por comprensión

- ▶ También se pueden crear **conjuntos por comprensión** usando la misma sintaxis de las **expresiones generadoras** y las **listas por comprensión**, pero esta vez encerrando la expresión entre llaves.
- ▶ Su sintaxis es:

```
<conj_comp> ::= { <expresión> (for <identificador> in <secuencia> [if <condición>])+ }
```

- ▶ Por ejemplo:

```
>>> {x ** 2 for x in [1, 2, 3]}  
{1, 4, 9}
```

- ▶ El resultado es directamente un valor de tipo `set`, no un iterador.

- ▶ Los conjuntos por comprensión, al igual que las expresiones generadoras y las listas por comprensión, **determinan su propio ámbito**.
- ▶ Ese ámbito abarca todo el conjunto por comprensión, de principio a fin.
- ▶ Los identificadores que aparecen en la cláusula **for** se consideran *variables ligadas* en el conjunto por comprensión.
- ▶ Esos identificadores se van ligando, uno a uno, a cada elemento de la secuencia indicada en la cláusula **in**.

► Como son variables ligadas, cumplen estas dos propiedades:

1. Se pueden renombrar (siempre de forma consistente) sin que el conjunto por comprensión cambie su significado.

Por ejemplo, los dos conjuntos por comprensión siguientes son equivalentes, puesto que producen el mismo resultado:

```
{x for x in (1, 2, 3)}
```

```
{y for y in (1, 2, 3)}
```

2. No se pueden usar fuera del conjunto por comprensión, ya que estarían fuera de su ámbito y no serían visibles.

Por ejemplo, lo siguiente daría un error de nombre:

```
>>> e = {x for x in (1, 2, 3)}  
>>> x      # Intento acceder a la 'x' del conjunto por comprensión  
Traceback (most recent call last):  
  File "<stdin>", line 1, in <module>  
NameError: name 'x' is not defined
```

Operaciones

- \underline{s} y \underline{o} son conjuntos, y \underline{x} es un valor cualquiera:

Operación	Resultado
<code>len(s)</code>	Número de elementos de \underline{s} (su cardinalidad)
<code>x in s</code>	<code>True</code> si \underline{x} pertenece a \underline{s}
<code>x not in s</code>	<code>True</code> si \underline{x} no pertenece a \underline{s}
<code>s.isdisjoint(o)</code>	<code>True</code> si \underline{s} no tiene ningún elemento en común con \underline{o}
<code>s.issubset(o)</code>	<code>True</code> si \underline{s} es un subconjunto de \underline{o}
<code>s <= o</code>	
<code>s < o</code>	<code>True</code> si \underline{s} es un subconjunto propio de \underline{o}
<code>s.issuperset(o)</code>	<code>True</code> si \underline{s} es un superconjunto de \underline{o}
<code>s >= o</code>	
<code>s > o</code>	<code>True</code> si \underline{s} es un superconjunto propio de \underline{o}

Operación	Resultado
<code>s.union(o)</code> <code>s o</code>	Unión de <u>s</u> y <u>o</u> ($s \cup o$)
<code>s.intersection(o)</code> <code>s & o</code>	Intersección de <u>s</u> y <u>o</u> ($s \cap o$)
<code>s.difference(o)</code> <code>s - o</code>	Diferencia de <u>s</u> y <u>o</u> ($s \setminus o$)
<code>s.symmetric_difference(o)</code> <code>s ^ o</code>	Diferencia simétrica de <u>s</u> y <u>o</u> ($s \Delta o$)
<code>s.copy()</code>	Devuelve una copia superficial de <u>s</u>

- ▶ Tanto `set` como `frozenset` admiten **comparaciones entre conjuntos**.
- ▶ Suponiendo que `a` y `b` son conjuntos:
 - `a == b` si y sólo si cada elemento de `a` pertenece también a `b`, y viceversa; es decir, si cada uno es un subconjunto del otro.
 - `a <= b` si y sólo si `a` es un *subconjunto* de `b` (es decir, si cada elemento de `a` está también en `b`).
 - `a < b` si y sólo si `a` es un *subconjunto propio* de `b` (es decir, si `a` es un subconjunto de `b`, pero no es igual a `b`).
 - `a >= b` si y sólo si `a` es un *superconjunto* de `b` (es decir, si cada elemento de `b` está también en `a`).
 - `a > b` si y sólo si `a` es un *superconjunto propio* de `b` (es decir, si `a` es un superconjunto de `b`, pero no es igual a `b`).

Operaciones sobre conjuntos mutables

- Estas tablas sólo se aplica a conjuntos mutables (o sea, al tipo `set` y no al `frozenset`):

Operación	Resultado
<code>s.update(o)</code> <code>s = o</code>	Actualiza <u>s</u> añadiendo los elementos de <u>o</u>
<code>s.intersection_update(o)</code> <code>s &= o</code>	Actualiza <u>s</u> manteniendo sólo los elementos que están en <u>s</u> y <u>o</u>
<code>s.difference_update(o)</code> <code>s -= o</code>	Actualiza <u>s</u> eliminando los elementos que están en <u>o</u>
<code>s.symmetric_difference_update(o)</code> <code>s ^= o</code>	Actualiza <u>s</u> manteniendo sólo los elementos que están en <u>s</u> y <u>o</u> pero no en ambos

Operación	Resultado
<code>s.add(x)</code>	Añade <u>x</u> a <u>s</u>
<code>s.remove(x)</code>	Elimina <u>x</u> de <u>s</u> (produce <code>KeyError</code> si <u>x</u> no está en <u>s</u>)
<code>s.discard(x)</code>	Elimina <u>x</u> de <u>s</u> si está en <u>s</u>
<code>s.pop()</code>	Elimina y devuelve un valor cualquiera de <u>s</u> (produce <code>KeyError</code> si <u>s</u> está vacío)
<code>s.clear()</code>	Elimina todos los elementos de <u>s</u>

Recorrido de conjuntos

- ▶ Como cualquier otro dato iterable, los conjuntos se pueden recorrer usando **iteradores**.
- ▶ El **orden** en el que se recorren los elementos del conjunto **no está determinado de antemano**, es decir, que el iterador puede entregar los elementos del conjunto **en cualquier orden**:
- ▶ Con un iterador:

```
>>> s = {3, 1, 2}
>>> it = iter(s)
>>> next(it)
1
>>> next(it)
2
>>> next(it)
3
```

- ▶ Con un bucle **for**:

```
>>> s = {3, 1, 2}
>>> for e in s:
...     print(e)
...
1
2
3
```

- ▶ Aunque, a la vista de este ejemplo, pudiera parecer que el conjunto siempre se va a recorrer como si estuviese ordenado, **no hay que confiar nunca** en que eso se vaya a cumplir siempre.

3.2. Diccionarios (`dict`)

3.2.1 Diccionarios por comprensión

3.2.2 Operaciones

3.2.3 Recorrido de diccionarios

Diccionarios (dict)

- ▶ Un **diccionario** es una colección que almacena *correspondencias* (o *asociaciones*) entre valores.
- ▶ Por tanto, **los elementos de un diccionario son parejas de valores llamados *clave* y *valor***, y lo que hace el diccionario es almacenar las *claves* y el *valor* que le corresponde a cada clave.
- ▶ Además, los elementos de un diccionario son datos mutables y, por tanto, los diccionarios también son **mutables**.
- ▶ En consecuencia, los diccionarios **NO** son *hashables*.

- Los diccionarios se pueden crear:
 - Con una pareja de llaves:

```
{}
```

que representan el **diccionario vacío**.

- Encerrando entre llaves una lista de parejas `<clave>: <valor>` separadas por comas:

```
{'juan': 4098, 'pepe': 4127}
```

Esa es precisamente la **forma normal** de un diccionario y, por tanto, la que se usa cuando se visualiza desde el intérprete o se imprime con `print`.

- Usando la función `dict`.

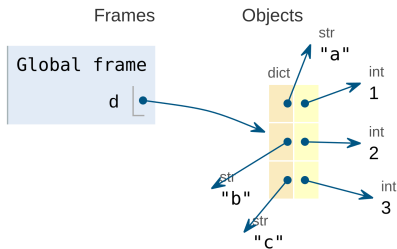
► Por ejemplo:

```
>>> v1 = {}                                # diccionario vacío
>>> v2 = dict()                            # también diccionario vacío
>>> v1 == v2
True                                       # son iguales
>>> a = {'uno': 1, 'dos': 2, 'tres': 3}    # literal
>>> b = dict(uno=1, dos=2, tres=3)         # argumentos con nombre
>>> c = dict([('dos', 2), ('uno', 1), ('tres', 3)]) # lista de tuplas
>>> d = dict({'tres': 3, 'uno': 1, 'dos': 2}) # crea una copia
>>> e = dict(zip(['uno', 'dos', 'tres'], [1, 2, 3])) # con dos iterables
>>> a == b and b == c and c == d and d == e # todos son iguales
True
```

- ▶ En memoria, los diccionarios se almacenan como **tablas** de dos columnas, la clave y el valor.
- ▶ Por ejemplo, el siguiente diccionario:

```
d = {'a': 1, 'b': 2, 'c': 3}
```

se almacenaría de la siguiente forma según lo representa la herramienta Pythontutor:



Diccionario almacenado en memoria

- ▶ Las **claves** de un diccionario deben cumplir dos **restricciones**:
 1. Deben ser **únicas** en ese diccionario.
 2. Deben ser **hashables**.

Claves únicas

- ▶ En un diccionario dado, **cada clave sólo puede asociarse con un único valor**.
- ▶ Por tanto, **en un diccionario no puede haber claves repetidas**, es decir, que no puede haber dos elementos distintos con la misma clave.
- ▶ Esto es así porque los elementos de un diccionario se identifican mediante su clave.

Así que, para acceder a un elemento dentro de un diccionario, debemos indicar la clave del elemento.

- ▶ Los tipos numéricos que se usen como claves obedecen las reglas normales de comparación numérica.
- ▶ Por tanto, si dos números son considerados iguales (como 1 y 1.0) entonces se consideran la misma clave dentro del diccionario.

- Si se intenta crear un diccionario con claves repetidas, sólo se almacenará uno de los elementos que tengan la misma clave (los demás se ignoran):

```
>>> a = {'perro': 'dog', 'gato': 'cat', 'perro': 'doggy'}  
>>> a  
{'perro': 'doggy', 'gato': 'cat'}
```

Como se ve, la clave `'perro'` está repetida y, por tanto, sólo se almacena uno de los dos elementos con clave repetida, que siempre es el último que se va a insertar en el diccionario.

En este caso, se almacena el elemento `'perro': 'doggy'` y se ignora el `'perro': 'dog'`.

Claves *hashables*

- ▶ Por otra parte, las **claves** de un diccionario deben ser datos ***hashables***.
- ▶ Por tanto, no se pueden usar como clave una lista, un conjunto `set`, otro diccionario o cualquier otro dato mutable.
- ▶ Si se intenta crear un diccionario con una clave no *hashable*, se produce un error `TypeError`:

```
>>> {[1, 2]: 'a', [3, 4]: 'b'}           # Las listas no son hashables
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unhashable type: 'list'
>>> {[1, 2]: 'a', {3, 4}: 'b'}          # Los conjuntos set tampoco
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unhashable type: 'set'
```

- ▶ En cambio, sí se puede usar un `frozenset`, al ser *hashable*:

```
>>> {frozenset({1, 2}): 'a', frozenset({3, 4}): 'b'}
{frozenset({1, 2}): 'a', frozenset({3, 4}): 'b'}
```

- ▶ Desde la versión 3.7 de Python, los elementos dentro de un diccionario se almacenan en **el orden en el que se van insertando dentro del diccionario**, aunque ese orden sólo tiene importancia en determinadas situaciones concretas.
- ▶ Dos diccionarios se consideran **iguales** si ambos contienen los mismos elementos, es decir, si tienen las mismas parejas *<clave>: <valor>*, sin importar el orden en el que aparezcan los elementos en el diccionario:

```
>>> {'a': 1, 'b': 2} == {'b': 2, 'a': 1}
True
```

- ▶ Para **acceder a un elemento** del diccionario se usa una sintaxis idéntica a la de la **indexación**, salvo que, en este caso, en lugar de usar el índice o posición del elemento, se usa la clave:

```
>>> a = {'perro': 'dog', 'gato': 'cat'}  
>>> a['perro']  
'dog'
```

- ▶ Si se intenta acceder a un elemento usando una clave que no existe, se lanza una excepción de tipo **KeyError**:

```
>>> a['caballo']  
Traceback (most recent call last):  
File "<stdin>", line 1, in <module>  
KeyError: 'caballo'
```

Diccionarios por comprensión

- ▶ También se pueden crear **diccionarios por comprensión** usando una sintaxis análoga a la de los **conjuntos por comprensión** (encerrando la expresión entre llaves), pero de forma que los elementos estén formados por parejas de clave y valor separados por `:`.
- ▶ Su sintaxis es:

```
<dicc_comp> ::= {<clave>:<valor> (for <identificador> in <secuencia> [if <condición>])+}
```

donde `<clave>` debe ser una expresión que devuelva valores *hashables*, y `<valor>` puede ser una expresión cualquiera.

- ▶ Por ejemplo:

```
>>> {x: x ** 2 for x in [1, 2, 3]}  
{1: 1, 2: 4, 3: 9}
```

devuelve el diccionario que asocia a cada número 1, 2 y 3 con su correspondiente cuadrado.

- ▶ El resultado es directamente un valor de tipo `dict`, no un iterador.

- ▶ Los diccionarios por comprensión, al igual que los conjuntos por comprensión, las expresiones generadoras y las listas por comprensión, **determinan su propio ámbito**.
- ▶ Ese ámbito abarca todo el diccionario por comprensión, de principio a fin.
- ▶ Los identificadores que aparecen en la cláusula **for** se consideran *variables ligadas* en el diccionario por comprensión.
- ▶ Esos identificadores se van ligando, uno a uno, a cada elemento de la secuencia indicada en la cláusula **in**.

► Como son variables ligadas, cumplen estas dos propiedades:

1. Se pueden renombrar (siempre de forma consistente) sin que el diccionario por comprensión cambie su significado.

Por ejemplo, los dos diccionarios por comprensión siguientes son equivalentes, puesto que producen el mismo resultado:

```
{x: x ** 2 for x in (1, 2, 3)}
```

```
{y: y ** 2 for y in (1, 2, 3)}
```

2. No se pueden usar fuera del diccionario por comprensión, ya que estarían fuera de su ámbito y no serían visibles.

Por ejemplo, lo siguiente daría un error de nombre:

```
>>> e = {x: x ** 2 for x in (1, 2, 3)}  
>>> x      # Intento acceder a la 'x' del diccionario por comprensión  
Traceback (most recent call last):  
  File "<stdin>", line 1, in <module>  
NameError: name 'x' is not defined
```

Operaciones

- d y o son diccionarios, c es una clave válida y v es un valor cualquiera:

Operación	Resultado
<code>d[c]</code>	Devuelve el valor asociado a <u>c</u> en <u>d</u> (lanza <code>KeyError</code> si <u>c</u> no está en <u>d</u>)
<code>d[c] = v</code>	Asocia a la clave <u>c</u> el valor <u>v</u> en <u>d</u> (crea el elemento dentro de <u>d</u> si la clave <u>c</u> no estaba ya en <u>d</u>)
<code>del d[c]</code>	Borra de <u>d</u> el elemento cuya clave es <u>c</u> (lanza <code>KeyError</code> si <u>c</u> no está en <u>d</u>)
<code>len(d)</code>	Número de elementos de <u>d</u>
<code>c in d</code>	<code>True</code> si <u>d</u> contiene un elemento con clave <u>c</u>
<code>c not in d</code>	<code>True</code> si <u>d</u> no contiene un elemento con clave <u>c</u>
<code>d.clear()</code>	Elimina todos los elementos de <u>d</u>
<code>d.copy()</code>	Devuelve una copia superficial de <u>d</u>

Operación	Resultado
<code>d.get(c[, def])</code>	Si la clave <u>c</u> está en <u>d</u> , devuelve <u>d[c]</u> ; si no está, devuelve <u>def</u> , que por defecto es <code>None</code>
<code>d.pop(c[, def])</code>	Si la clave <u>c</u> está en <u>d</u> , devuelve <u>d[c]</u> y elimina de <u>d</u> el elemento con clave <u>c</u> ; si no está, devuelve <u>def</u> (si no se pasa <u>def</u> y la clave <u>c</u> no está en <u>d</u> , lanza un <code>KeyError</code>)
<code>d.popitem()</code>	Selecciona un elemento de <u>d</u> siguiendo un orden LIFO, lo elimina de <u>d</u> y lo devuelve en forma de tupla (<u>clave</u> , <u>valor</u>) (lanza un <code>KeyError</code> si <u>d</u> está vacío)
<code>d.setdefault(c[, def])</code>	Si la clave <u>c</u> está en <u>d</u> , devuelve <u>d[c]</u> ; si no está, inserta en <u>d</u> un elemento con clave <u>c</u> y valor <u>def</u> , y devuelve <u>def</u> (que por defecto es <code>None</code>)
<code>d.update(o)</code>	Actualiza <u>d</u> con las parejas (<u>clave</u> , <u>valor</u>) de <u>o</u> , sobrescribiendo las claves ya existentes en <u>d</u> , y devuelve <code>None</code>

Recorrido de diccionarios

- ▶ Como cualquier otro dato iterable, los diccionarios se pueden recorrer usando iteradores.
- ▶ El orden en el que se recorren los elementos del diccionario es el orden en el que están almacenados los elementos dentro del diccionario que, como ya sabemos, desde la versión 3.7 de Python coincide con el orden en el que se han ido insertando los elementos en el diccionario.
- ▶ Los iteradores creados sobre un diccionario, en realidad, recorren sus **claves**:
- ▶ Con un iterador:
- ▶ Con un bucle **for**:

```
>>> d = {'a': 1, 'b': 2}
>>> it = iter(d)
>>> next(it)
'a'
>>> next(it)
'b'
```

```
>>> d = {'a': 1, 'b': 2}
>>> for k in d:
...     print(k)
...
a
b
```

- ▶ Si, además de acceder a las **claves**, necesitamos también acceder a los **valores** del diccionario mientras lo recorremos, podemos:
 - Acceder al valor a partir de la clave usando **indexación**:

```
>>> for k in d:  
...     print(k, d[k])  
...  
a 1  
b 2
```

- Usar el método `items` sobre el diccionario (el cual devuelve un objeto que, al iterar sobre él, genera una secuencia de tuplas (`<clave>`, `<valor>`)), y combinarlo con el **desempaquetado de tuplas**:

```
>>> d.items()  
dict_items([('a', 1), ('b', 2)])  
>>> for k, v in d.items():  
...     print(k, v)  
...  
a 1  
b 2
```

- ▶ Otros métodos útiles para recorrer un diccionario son `keys` y `values`.
- ▶ `keys` devuelve un objeto que, al iterar sobre él, va generando las **claves** del diccionario sobre el que se invoca:
- ▶ `values` devuelve un objeto que, al iterar sobre él, va generando los **valores** del diccionario sobre el que se invoca:

```
>>> d.keys()
dict_keys(['a', 'b'])
>>> for k in d.keys():
...     print(k)
...
a
b
```

```
>>> d.values()
dict_values([1, 2])
>>> for v in d.values():
...     print(v)
...
1
2
```

- ▶ En la práctica, no resulta muy útil usar `keys`, ya que se puede hacer lo mismo recorriendo directamente el propio diccionario, como ya sabemos:

```
>>> for k in d:
...     print(k)
...
a
b
```

3.3. Documentos XML

3.3.1 Acceso

3.3.2 Modificación

Documentos XML

- ▶ Los documentos XML se pueden considerar **datos estructurados en forma de árbol** (es decir, con una estructura **jerárquica** y, por tanto, **no secuencial**).
- ▶ Por ejemplo, supongamos el siguiente documento XML:

```
<?xml version="1.0"?>
<raiz>
  <alumno numero="111">
    <dni>12312312A</dni>
    <nombre>
      <propio>Juan</propio>
      <apellidos>García González</apellidos>
    </nombre>
    <telefono>666555444</telefono>
    <nota>7</nota>
  </alumno>
  <alumno numero="222">
    <dni>44455566B</dni>
    <nombre>
      <propio>María</propio>
      <apellidos>Pérez Rodríguez</apellidos>
    </nombre>
    <telefono>696969696</telefono>
    <nota>9</nota>
  </alumno>
  <madre>
    <dni>22222222C</dni>
  </madre>
</raiz>
```

► Podemos encontrarnos lo siguiente:

- Lo que hay entre ángulos (como `<raiz>`) es una *etiqueta*.
- Lo que hay entre dos etiquetas (como el `7` en `<nota>7</nota>`) es el *texto* de la etiqueta.
- Lo que hay dentro de la etiqueta (como `numero="111"`) es un *atributo* de la etiqueta.

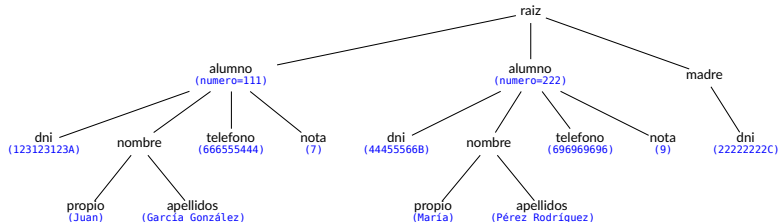
`numero="111"` es un atributo de la primera etiqueta `<alumno>`.

`numero` es el *nombre* del atributo.

`111` es el *valor* del atributo `numero`.

- Puede haber etiquetas con uno o varios atributos y etiquetas sin atributos.

- Ese documento representaría la siguiente estructura jerárquica:



- Se observa que:
 - Cada nodo representa una etiqueta del documento XML.
 - Si una etiqueta contiene a otra, su correspondiente nodo en el árbol tendrá un hijo que representa a la etiqueta que contiene.
 - Los hijos de un nodo están ordenados por la posición que ocupan dentro de su etiqueta padre.

Acceso

- ▶ El módulo `xml.etree.ElementTree` (documentado en <https://docs.python.org/3/library/xml.etree.elementtree.html>) implementa una interfaz sencilla y eficiente para interpretar y crear datos XML.
- ▶ Para importar los datos de un archivo XML, podemos hacer:

```
import xml.etree.ElementTree as ET
arbol = ET.parse('archivo.xml')
raiz = arbol.getroot()
```

- ▶ Si los datos XML se encuentran ya en una cadena, se puede hacer directamente:

```
raiz = ET.fromstring(datos_en_una_cadena)
```

- Los nodos del árbol se representan internamente mediante objetos de tipo `Element`, los cuales disponen de ciertos atributos y responden a ciertos métodos.

¡Cuidado! Aquí, cuando hablamos de *atributos*, nos referimos a información que contiene el objeto (una cualidad del objeto según el *paradigma orientado a objetos*) y a la cual se puede acceder usando el operador punto (`.`), no a los atributos que pueda tener una etiqueta según consten en el documento XML.

- Los objetos de tipo `Element` disponen de los siguientes atributos:
 - `tag`: una cadena que representa a la etiqueta del nodo (por ejemplo: si la etiqueta es `<alumno>`, entonces `tag` contendrá `'alumno'`).
 - `attrib`: un diccionario que representa a los atributos de esa etiqueta en el documento XML.
 - `text`: una cadena que representa el contenido del nodo, es decir, el texto que hay entre `<etiqueta>` y `</etiqueta>`.

- Por ejemplo, si tenemos la siguiente etiqueta en el documento XML:

```
<telefono tipo="movil">666555444</telefono>
```

y la variable `nodo` contiene el nodo (es decir, el objeto `Element`) que representa a dicha etiqueta en el árbol, entonces:

- `nodo.tag` valdrá `'telefono'`.
- `nodo.attrib` valdrá `{'tipo': 'movil'}`.
- `nodo.text` valdrá `'666555444'`.

- ▶ En nuestro caso, `raiz` es un objeto de tipo `Element` que, además, representa al nodo raíz del árbol XML.
- ▶ Por tanto, tendríamos lo siguiente:

```
>>> raiz.tag
'raiz'
>>> raiz.attrib
{}
>>> raiz.text
'\n'                                     # ¿Por qué?
```

- ▶ Los objetos `Element` son **iterables**. Por ejemplo, el nodo raíz tiene **nodos hijos** (nodos que «cuelgan» directamente del nodo raíz) sobre los cuales se puede iterar desde el objeto `raiz`:

```
>>> for hijo in raiz:
...     print(hijo.tag, hijo.attrib)
...
alumno {'numero': '111'}
alumno {'numero': '222'}
madre {}
```

- Los hijos están anidados, y podemos acceder a nodos concretos a través de su índice (es decir, que los objetos `Element` son **indexables**):

```
>>> raiz[0]          # el primer hijo directo de raiz
<Element 'alumno' at 0x7f929c29cf90>
>>> raiz[0][2]       # el tercer hijo directo del primer hijo directo de raiz
<Element 'telefono' at 0x7f929c29d180>
>>> raiz[0][2].text
'666555444'
```

- ▶ Los objetos de tipo `Element` disponen de métodos útiles para iterar recursivamente sobre todos los subárboles situados debajo de él (sus hijos, los hijos de sus hijos, y así sucesivamente).
- ▶ Por ejemplo, el método `iter` devuelve un **iterador** que recorre todos los nodos del árbol desde el nodo actual (el nodo sobre el que se invoca al método) en un orden *primero en profundidad*.
- ▶ Eso quiere decir que va visitando los nodos en el mismo orden en el que se encuentran escritos en el documento XML, incluyendo el propio nodo sobre el que se invoca.

► Por ejemplo:

```
>>> for nodo in raiz.iter():  
...     print(nodo.tag)  
...  
raiz  
alumno  
dni  
nombre  
propio  
apellidos  
telefono  
nota  
alumno  
dni  
nombre  
propio  
apellidos  
telefono  
nota  
madre  
dni
```

- Si se le pasa una etiqueta como argumento, devolverá únicamente los nodos que tengan esa etiqueta:

```
>>> for dni in raiz.iter('dni'):  
...     print(dni.text)  
...  
12312312A  
44455566B  
22222222C
```

- El método `findall` devuelve una lista con los nodos que tengan una cierta etiqueta y que sean hijos directos del nodo sobre el que se invoca.

Puede devolver una lista vacía si no hay nodos que cumplan la condición.

- El método `find` devuelve el primer hijo directo del nodo sobre el que se invoca, siempre que tenga una cierta etiqueta indicada como argumento.

Puede devolver `None` si el nodo no tiene ningún hijo con esa etiqueta.

- El método `get` devuelve el valor de algún atributo de la etiqueta asociada a ese nodo:

```
>>> for alumno in raiz.findall('alumno'):
...     numero = alumno.get('numero')
...     dni = alumno.find('dni').text
...     print(numero, dni)
...
111 12312312A
222 44455566B
```


- Si la etiqueta no tiene el atributo indicado en el argumento de `get`, éste devuelve `None` o el valor que se haya indicado en el segundo argumento:

```
>>> alumno = raiz[0]
>>> alumno.get('numero')
111
>>> alumno.get('altura')
None
>>> alumno.get('altura', 0)
0
>>> alumno.get('numero', 0)
111
```

- También disponemos de la función `dump` que devuelve la cadena correspondiente al nodo que se le pase como argumento:

```
>>> ET.dump(raiz[0][2])
<telefono>666555444</telefono>
```

- Para una especificación más sofisticada de los elementos a encontrar, se pueden usar las expresiones **XPath** con los métodos **find** y **findall**:

Sintaxis	Significado
etiqueta	Selecciona todos los nodos hijo con la etiqueta etiqueta . Por ejemplo, pepe selecciona todos los nodos hijo llamados pepe , y pepe/juan selecciona todos los nietos llamados juan en todos los hijos llamados pepe .
*	Selecciona todos los nodos hijo inmediatos. Por ejemplo, */pepe selecciona todos los nietos llamados pepe .
.	Selecciona el nodo actual. Se usa, sobre todo, al principio de la ruta para indicar que es una ruta relativa.
//	Selecciona todos los subnodos en cualquier nivel por debajo de un nodo. Por ejemplo, ./pepe selecciona todos los nodos pepe que haya en todo el árbol.
..	Selecciona el nodo padre. Devuelve None si se intenta acceder a un ancestro del nodo de inicio (aquel sobre el que se ha llamado al método find).

► Continuación de las expresiones **XPath**:

Sintaxis	Significado
<code>[@atrib]</code>	Selecciona todos los nodos que tienen el atributo atrib .
<code>[@atrib= 'valor']</code>	Selecciona todos los nodos que tienen el valor valor en el atributo atrib . El valor no puede contener apóstrofes.
<code>[@atrib!= 'valor']</code>	Selecciona todos los nodos que tienen el valor valor en el atributo atrib . El valor no puede contener apóstrofes.
<code>[etiqueta]</code>	Selecciona todos los nodos que tienen un hijo inmediato llamado etiqueta .
<code>[posición]</code>	Selecciona todos los nodos situados en cierta posición . Ésta puede ser un entero (1 es la primera posición), la expresión <code>last()</code> (la última posición) o una posición relativa a la última posición (por ejemplo, <code>last() - 1</code>).

Ejemplos

```
# Los nodos de nivel más alto:
>>> raiz.findall(".")
[<Element 'raiz' at 0x7f929c29cf40>]

# Los nietos 'dni' de los hijos 'alumno' de los nodos de nivel más alto:
>>> raiz.findall("./alumno/dni")
[<Element 'dni' at 0x7f929c29d040>,
 <Element 'dni' at 0x7f929c29d270>]

# Lo de antes equivale a hacer (porque el nodo actual es el raíz):
>>> raiz.findall("alumno/dni")
[<Element 'dni' at 0x7f929c29d040>,
 <Element 'dni' at 0x7f929c29d270>]

# Los nodos con numero='111' que tienen un hijo 'dni':
>>> raiz.findall("./dni/..[@numero='111']")
[<Element 'alumno' at 0x7f929c29cf90>]

# Antes de // hay que poner algo que indique el nodo debajo del
# cual se va a buscar:
>>> raiz.findall("//dni/..[@numero='111']")
SyntaxError: cannot use absolute path on element
```

Ejemplos

```
# Todos los nodos 'dni' del árbol completo:
>>> raiz.findall('./dni')
[<Element 'dni' at 0x7f547a0e9950>,
 <Element 'dni' at 0x7f547a0e9b80>,
 <Element 'dni' at 0x7f547a0e9e00>]

# Sólo los DNIs que estén por debajo de un nodo 'madre'
# en cualquier nivel:
>>> raiz.findall('madre/dni')
[<Element 'dni' at 0x7f547a0e9e00>]

# Los nodos 'dni' que son hijos de los nodos con numero='111':
>>> raiz.findall(".*[@numero='111']/dni")
[<Element 'dni' at 0x7f929c29d040>]

# Los nodos 'alumno' que son hijos segundos de sus padres:
>>> raiz.findall("./alumno[2]")
[<Element 'alumno' at 0x7f929c29d220>]

# Los nodos 'alumno' hijos directos del actual que tienen un hijo 'nota':
>>> raiz.findall("./alumno[nota]")
[<Element 'alumno' at 0x7f929c29cf90>,
 <Element 'alumno' at 0x7f929c29d220>]

# Lo de antes equivale a hacer (porque el nodo actual es el raíz):
>>> raiz.findall("alumno[nota]")
[<Element 'alumno' at 0x7f929c29cf90>,
 <Element 'alumno' at 0x7f929c29d220>]
```

Modificación

- ▶ `ElementTree` proporciona una forma sencilla de crear documentos XML y escribirlos en archivos.
- ▶ Para ello, usamos el método `write`.
- ▶ Una vez creado, un objeto `Element` puede manipularse directamente:
 - Cambiando los atributos del objeto, como `text` o `attrib`.
 - Cambiando los atributos de la etiqueta a la que representa el objeto, con el método `set`.
 - Añadiendo nuevos hijos con los métodos `append` o `insert`.
 - Eliminando hijos con el método `remove`.

- Por ejemplo, supongamos que queremos sumarle 1 a la **nota** de cada alumno y añadir un atributo **modificado** a la etiqueta **nota**:

```
>>> for nota in raiz.iter('nota'):
...     nueva_nota = int(nota.text) + 1
...     nota.text = str(nueva_nota)
...     nota.set('modificado', 'si')
...
>>> arbol.write('salida.xml')
```

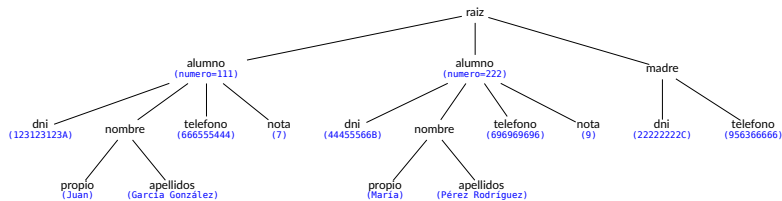
- Nuestro XML tendría ahora el siguiente aspecto:

```
<?xml version="1.0"?>
<raiz>
  <alumno numero="111">
    <dni>12312312A</dni>
    <nombre>
      <propio>Juan</propio>
      <apellidos>García González</apellidos>
    </nombre>
    <telefono>666555444</telefono>
    <nota modificado="si">8</nota>
  </alumno>
  <alumno numero="222">
    <dni>44455566B</dni>
    <nombre>
      <propio>Maria</propio>
      <apellidos>Pérez Rodríguez</apellidos>
    </nombre>
    <telefono>696969696</telefono>
    <nota modificado="si">10</nota>
  </alumno>
  <madre>
    <dni>22222222C</dni>
  </madre>
</raiz>
```


- Podemos insertar nuevos nodos hijos de un determinado nodo con los métodos `append` o `insert`, como si el nodo fuese una lista.
- Para ello, primero normalmente crearemos el nodo que vamos a insertar usando `Element(etiqueta)`.
- Por ejemplo, añadimos el teléfono a la única madre que tenemos en el documento XML:

```
telefono = ET.Element('telefono')
telefono.text = '956366666'
madre = raiz.find('madre')
madre.append(telefono)
```

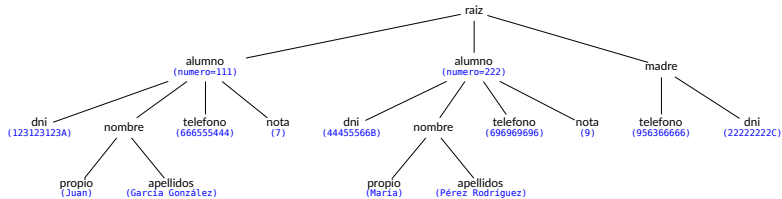
- Tras estas operaciones, ahora tendríamos:



- Podríamos haber usado `insert` en lugar de `append` para cambiar la posición donde situar el nodo hijo.
- Por ejemplo, si queremos situar el teléfono antes que el DNI, podríamos hacer:

```
telefono = ET.Element('telefono')  
telefono.text = '956366666'  
madre = raiz.find('madre')  
madre.insert(0, telefono)
```

- Tras estas operaciones, ahora tendríamos:



- Podemos eliminar elementos con el método `remove`. Por ejemplo, supongamos que queremos eliminar todos los alumnos con una `nota` inferior a 9:

```
>>> for alumno in raiz.findall('alumno'):
...     # se usa findall para que no afecte el borrado durante el recorrido
...     nota = int(alumno.find('nota').text)
...     if nota < 9:
...         raiz.remove(alumno)
...
>>> arbol.write('salida.xml')
```

- Tener en cuenta que la modificación concurrente mientras se hace una iteración puede dar problemas, lo mismo que ocurre cuando se modifican listas o diccionarios mientras se itera sobre ellos.
- Por ello, el ejemplo primero recoge todos los elementos con `findall` y sólo entonces itera sobre la lista que devuelve.
- Si usáramos `iter` en lugar de `findall` se podrían dar problemas debido a que `iter` va devolviendo perezosamente los nodos (es un iterador) y el conjunto de nodos que devuelve podría verse afectado por los borrados.

- Nuestro XML tendría ahora el siguiente aspecto:

```
<?xml version="1.0"?>
<raiz>
  <alumno numero="222">
    <dni>44455566B</dni>
    <nombre>
      <propio>María</propio>
      <apellidos>Pérez Rodríguez</apellidos>
    </nombre>
    <telefono>696969696</telefono>
    <nota modificado="si">10</nota>
  </alumno>
  <madre>
    <dni>22222222C</dni>
  </madre>
</raiz>
```

- Si lo que queremos es **mover** un nodo (es decir, cambiar un nodo de sitio), podemos combinar los efectos de `append` e `insert` con `remove`.
- Por ejemplo, si queremos mover la etiqueta `<madre>` dentro de la etiqueta `<alumno>`, podríamos hacer:

```
madre = raiz.find('madre')
raiz.remove(madre)
alumno = raiz.find('alumno')
alumno.append(madre)
```

- Nuestro XML tendría ahora el siguiente aspecto:

```
<?xml version="1.0"?>
<raiz>
  <alumno numero="222">
    <dni>44455566B</dni>
    <nombre>
      <propio>María</propio>
      <apellidos>Pérez Rodríguez</apellidos>
    </nombre>
    <telefono>696969696</telefono>
    <nota modificado="si">10</nota>
    <madre>
      <dni>22222222C</dni>
    </madre>
  </alumno>
</raiz>
```


- La función `SubElement` también proporciona una forma muy conveniente de crear sub-elementos de un elemento dado:

```
>>> a = ET.Element('a')
>>> b = ET.SubElement(a, 'b')
>>> c = ET.SubElement(a, 'c')
>>> d = ET.SubElement(c, 'd')
>>> ET.dump(a)
<a><b /><c><d /></c></a>
```

4. Bibliografía

Bibliografía

Python Software Foundation. n.d. "Sitio Web de Documentación de Python."
<https://docs.python.org/3>.