

Python for Data Science

IISER-Kolkata ML4HEP Pre-school Lecture Series
May 13, 2025 - Lecture 4

Subir Sarkar, SINP
subir.sarkar@cern.ch

<https://sarkar.web.cern.ch>

String

String

- A string in Python is a sequence of characters
- Strings are **immutable**

```
a = "Higgs boson"
b = 'aka, the God particle'
c = """ # span multiple lines
was discovered
at the Large Hadron Collider
10 years ago
"""

# quoted string
print("He asked, "Which is your favorite team?"")
print('He asked, "Which is your favorite team?"')
```

String

```
>>> import string # load the module in memory
>>> dir(string)
[... 'atof', 'atof_error', 'atoi',
'atoi_error', 'atol', 'atol_error',
'capitalize', 'capwords', 'center', 'count',
'digits', 'expandtabs', 'find', 'hexdigits',
'index', 'index_error', 'join', 'joinfields',
'letters', 'ljust', 'lower', 'lowercase',
'lstrip', 'maketrans', 'octdigits',
'printable', 'punctuation', 'replace',
'rfind', 'rindex', 'rjust', 'rsplit',
'rstrip', 'split', 'splitfields', 'strip',
'swapcase', 'translate', 'upper',
'uppercase', 'whitespace', 'zfill']
```

String object operations

```
>>> s = 'Hello World!' >>> s.startswith('Hello')
True
>>> s.isdigit()
False
>>> s.find('W')
6
>>> s.rfind('W')
6
>>> s.endswith('W')
False
>>> s.split(" ")
['Hello', 'World!']
>>> s.upper()
'HELLO WORLD!'
>>> s.swapcase()
'hELLO wORLD!'
>>> s.replace('World', 'Globe')
'Hello Globe!'
>>> s = 'Hello\tWorld!'
>>> s
'Hello\tWorld!'
>>> s.expandtabs()
'Hello    World!'
```

String operations

Convert a string

'1,2,3,4,5,6,7,8'

to

'1:2:3:4:5:6:7:8'

String operations : split & join

```
>>> astr = '1,2,3,4,5,6,7,8'
>>> list = astr.split(',')
>>> list
['1', '2', '3', '4', '5', '6', '7', '8']
```

```
>>> ':' . join(list)
'1:2:3:4:5:6:7:8'
```

all in one line

```
>>> ':' . join('1,2,3,4,5,6,7,8'.split(','))
'1:2:3:4:5:6:7:8'
```

String Formatting [C style]

```
>>> print("Today's price: %f" % 50.4625)
```

```
Today's price: 50.462500
```

```
>>> print("Today's price: %.2f" % 50.4625)
```

```
Today's price: 50.46
```

```
>>> print("Change since yesterday: %+ .2f"  
% 1.5)
```

```
Change since yesterday: +1.50
```

String Formatting [Modern]

```
>>> print("{:.3f}".format(math.pi))
```

```
3.142
```

```
>>> print("{0:d} - {0:x} - {0:o} - {0:b}".format(21))
```

```
21 - 15 - 25 - 10101
```

```
>>> s = "I prefer {0} over {1}".format('emacs', 'vi')
```

```
>>> s = "Lang: {lang}".format(lang="Python")
```

```
# defining format
```

```
>>> email_f = "Your email was {email}".format
```

```
>>> print(email_f(email="bob@example.com"))
```

Basic Input/Output

Basic I/O

```
>>> import sys
>>> dir(sys) # returns a list
[... '__stderr__', '__stdin__',
'__stdout__', 'maxsize',
'setrecursionlimit', 'settrace', 'stderr',
'stdin', 'stdout', 'subversion',
'version', 'version_info', 'warnoptions']

>>> type(sys.stdin)
<class '_io.TextIOWrapper'>

>>> type(sys.stdout)
<class '_io.TextIOWrapper'>

>>> type(sys.stderr)
<class '_io.TextIOWrapper'>
```

Basic I/O

- Redirect stdout to a file

```
>>> import sys
>>> f = open('test.txt', 'w')
>>> sys.stdout = f
>>> print('hello world!')
>>> f.close()
>>> sys.stdout = sys.__stdout__
>>> print('hello world!')
hello world!
```

Basic I/O

```
>>> data = sys.stdin.readlines()
```

```
hello
```

```
world!
```

```
how
```

```
do
```

```
you do?
```

```
# CTRL-D
```

```
>>> print ("Counted", len(data), "lines.")
```

```
Counted 6 lines.
```

```
>>> print (sys.stdin.read().upper())
```

```
Hello World! # press CTRL-D
```

```
HELLO WORLD!
```

built-in function input

```
>>> str = input("Enter your input: ")
```

```
Enter your input: [x*5 for x in  
range(2,10,2)]
```

```
>>> print("evaluated input:", eval(str))
```

```
Received input is: [10, 20, 30, 40]
```

File I/O

player, country, testrun, odi run

Gavaskar, India, 10122, 3092

Border, Aus, 11174, 6524

Waugh, Aus, 10927, 7569

Lara, WI, 11953, 10405

Chanderpaul, WI, 11219, 8778

Jayawardene, SL, 11319, 11549

Tendulkar, India, 15921, 18426

Ponting, Aus, 13378, 13704

Kallis, SA, 13289, 11574

Dravid, India, 13288, 10889

Sangakkara, SL, 11151, 12548

File I/O

```
f = open('players.csv', 'r') # r -> read
for line in f:
    line = line.strip()
    fields = line.split(",")
    if fields[1] == 'India':
        print(line)
f.close()
```

File I/O

```
f = open('players.csv', 'r') # r -> read
for line in f:
    line = line.strip()
    fields = line.split(",")
    if fields[1] == 'India':
        print(line)
f.close()
```

- **what happens if**

- the file does not exist?
- there is a comment line?
- some lines are empty?
- some rows do not have all the required fields?

File I/O – better approach

```
with open('players.csv', 'r') as f:
    for line in f:
        line = line.strip()
        if (len(line) == 0): continue
        if line.find('#'):
            continue
        fields = line.split(",")
        if len(fields) != 4: continue
        if fields[1] == 'India':
            print(line)
```

file will be automatically closed

File Output

```
# Open a file
fo = open("foo.txt", "w")
fo.write("Python is great!!\n")

# Close opened file
fo.close()
```