第五章 - 字符串

张建章

阿里巴巴商学院

杭州师范大学

2024-09

字符串是**不可变的**（immutable），一旦创建，字符串中的字符无法直接修改。Python 字符串支持**多种操作和方法**，如字符串的分割（`split`）、替换（`replace`）、查找（`find`）和大小写转换（`upper`，`lower`）等。这些操作在处理文本数据时非常有效，能够简化对大规模文本的分析和处理。

字符串是一种不可变的字符序列，创建字符串的基本语法较为简单，主要通过以下几种方式实现：

1. 单引号或双引号创建字符串

使用单引号（`' '`）或双引号（`" "`）可以创建字符串。例如：

```
string1 = 'Hello, World'
string2 = "Hello, Python"
print(string1)  # 输出: Hello, World
print(string2)  # 输出: Hello, Python
```

Python 允许在字符串中使用单双引号，只要引号的形式匹配。

2. 多行字符串

使用三重引号（`''' '''` 或 `""" """`）可以创建多行字符串。适用于较长的文本或需要保留格式的文本内容。例如：

```
message = """This is a multi-line
string example."""
print(message)
```

输出将保留原始的换行格式。

3. `str()`

使用 `str()` 函数创建空字符串或将其他数据类型转换为字符串。

4. 字符串的不可变性：一旦字符串被创建，字符串中的字符无法被修改。这意味着尝试修改字符串中的某个字符会导致错误：

```
s = "hello"
s[0] = 'H'  # 报错: TypeError: 'str' object does not support item
↪  assignment
# 但可以通过创建新的字符串来变更其内容
new_s = 'H' + s[1:]
print(new_s)  # 输出: Hello
```

5. f-string 格式化：Python 提供了 f-string 格式化方式，允许在字符串中嵌入变量或表达式。例如：

```
name = "Alice"
age = 25
print(f"My name is {name} and I am {age} years old.")
```

f-string 不仅语法简洁，还支持嵌入复杂的表达式。

字符串的基本操作包括通过索引访问字符、使用切片提取子字符串、计算长度、进行成员资格检查、连接与重复字符串操作，以及需注意字符串的不可变性，无法直接修改其中的元素。

1. 索引

假设有一个字符串 `s = "Python"`，可以通过索引访问各个字符：

```
s = "Python"
print(s[0])   # 输出：P
print(s[1])   # 输出：y
print(s[-1])  # 输出：n  (使用负索引)
print(s[-2])  # 输出：o  (倒数第二个字符)
```

如果尝试访问超出字符串长度的索引（例如 `s[10]`），会引发 `IndexError` 错误。

2. 切片

一种从字符串中提取子字符串的方式，**语法同列表切片**。

```
s = "ABCDEFGHI"
print(s[2:7])    # 输出: CDEFG
print(s[:5])     # 输出: ABCDE (从索引 0 开始，步长为 1)
print(s[4:])     # 输出: EFGHI (从索引 4 开始直到结束)
print(s[::2])    # 输出: ACEGI (每隔一个字符提取)
```

负索引允许从字符串的末尾开始计数。例如：

```
print(s[-4:-1])    # 输出: FGHI (从倒数第4个字符开始提取到倒数第2个)
print(s[::-1])     # 输出: IHGFEDCBA (反转字符串)
```

3. 计算长度

计算字符串的长度可以通过内置函数 `len()` 完成。该函数接受一个字符串作为参数，并返回其中字符的总数，包括空格和标点符号。

```
# 定义一个字符串
feedback = "Customer service was excellent!"
# 计算字符串的长度
length = len(feedback)
# 输出结果
print(length)  # 输出: 31
```

4. 成员资格检查：通过 `in` 和 `not in` 运算符完成，用于检测子字符串是否存在于给定的字符串中。

```
feedback = "The product quality is excellent."
result = "excellent" in feedback
print(result)  # 输出: True
```

5. 字符串连接：可以通过使用 `+` 操作符，将多个字符串组合成一个新的字符串。

```
greeting = "Hello"
name = "Alice"
result = greeting + " " + name
print(result)  # 输出: Hello Alice
```

在这个例子中，`+` 操作符将两个字符串连接起来，生成新的字符串 `"Hello Alice"`。

另一种高效的方式是使用 `join()` 方法，将一个可迭代对象（如列表）中的元素连接为一个字符串，尤其是在处理大量字符串时更为节省内存：

```
words = ["Python", "is", "fun"]
result = " ".join(words)
print(result)  # 输出: Python is fun
```

6. 字符串重复字符串的重复可以使用 `*` 操作符实现，将一个字符串按指定次数重复。例如：

```
repeat_str = "ha " * 3
print(repeat_str)  # 输出: ha ha ha
```

这种操作常用于生成格式化的分隔符或模式，例如：

```
line = "=" * 10
print(line)  # 输出: ==========
```

1. 字符串拆分：`split()`

`split()`方法用于按照指定的分隔符将字符串拆分成子字符串列表。默认情况下，`split()`会按照空格分隔字符串。如果需要，可以通过传递参数指定不同的分隔符。

```
sentence = "Python is fun to learn"
words = sentence.split()  # 按空格拆分
print(words)  # 输出: ['Python', 'is', 'fun', 'to', 'learn']

# 使用指定分隔符拆分
sentence = "name,email,phone"
fields = sentence.split(',')
print(fields)  # 输出: ['name', 'email', 'phone']
```

2. 字符串合并：`join()`

`join()` 方法用于将一个可迭代对象（如列表或元组）中的元素通过指定的分隔符连接成一个字符串。`join()` 方法调用时应在分隔符字符串上调用，并传入需要合并的字符串列表。

```
words = ['Python', 'is', 'fun']
sentence = ' '.join(words)  # 使用空格连接
print(sentence)  # 输出: Python is fun

# 使用自定义分隔符
fields = ['name', 'email', 'phone']
csv_format = ','.join(fields)
print(csv_format)  # 输出: name,email,phone
```

这些操作在处理结构化文本数据（如 CSV 文件）或构建文本报告时非常有用。

3. 字符串查找：`find()`方法

`find()`方法用于在字符串中查找子字符串的索引位置。如果找到匹配的子字符串，返回其起始索引；否则返回`-1`。可以指定可选的起始和结束索引，限制查找的范围。

```
text = "Revenue for the year is estimated at $5 million."
position = text.find("estimated")
print(position)  # 输出: 25
```

在该示例中，`find()`方法返回子字符串`"estimated"`在字符串中的位置。

2. 字符串查找：`index()` 方法

字符串的 `index()` 方法用于查找子字符串在主字符串中的位置。其基本语法为：

```
str.index(sub[, start[, end]])
```

- `sub`：要搜索的子字符串。
- `start`：可选，搜索的起始位置。
- `end`：可选，搜索的结束位置。

如果找到该子字符串，`index()` 返回其在主字符串中的最低索引；若未找到，则抛出 `ValueError` 异常。以下是几个示例代码：

```python
sentence = "Hello, world!"
position = sentence.index("world")
print(position)  # 输出：7
```

`start` 和 `end` 参数含义同字符串切片：

```
# 使用起始参数
phrase = "Python is great. Python is versatile."
position = phrase.index("Python", 10)
print(position)  # 输出: 21
# 使用结束参数
phrase = "Python is great. Python is versatile."
position = phrase.index("le", 10, 35) # ValueError
position = phrase.index("le", 10, 36)
print(position)  # 输出: 34
```

`index()` 方法在处理字符串搜索时非常有效，尤其是在确定子字符串存在的情况下。

3. 字符串替换：`replace()` 方法

`replace()` 方法用于将字符串中的某个子字符串替换为另一个子字符串。它的基本语法是：

```
str.replace(old, new, count)
```

其中 `old` 是要替换的子字符串，`new` 是替换后的字符串，`count` 是可选参数，表示替换的次数。如果不指定 `count`，将替换所有出现的子字符串。

```
report = "The profit margin was low. The profit margin needs
↪  improvement."
new_report = report.replace("profit margin", "revenue")
print(new_report)
# 输出: The revenue was low. The revenue needs improvement.
```

在此示例中，`replace()` 方法将所有出现的 `"profit margin"` 替换为 `"revenue"`，生成了一个新的字符串。

4. 大小写转换

字符串的大小写转换可以通过以下几种常用的内置方法完成，包括`upper()`、`lower()`、`capitalize()`和`swapcase()`，这些方法在处理文本数据时非常有用，尤其是在标准化、数据清洗和文本分析的场景中。

`upper()`方法将字符串中的所有字母转换为大写：

```
text = "python is fun"
upper_text = text.upper()
print(upper_text)  # 输出: PYTHON IS FUN
```

`lower()`方法用于将字符串中的所有字母转换为小写：

```
text = "Hello, WORLD!"
lower_text = text.lower()
print(lower_text)  # 输出: hello, world!
```

`capitalize()` 方法将字符串的第一个字母转换为大写，其他字母转换为小写，适用于标题或句子的首字母格式化：

```
text = "python programming"
capitalized_text = text.capitalize()
print(capitalized_text)  # 输出: Python programming
```

`swapcase()` 方法将字符串中的大写字母转换为小写，小写字母转换为大写：

```
text = "PyThOn PrOgRaMmInG"
swapped_text = text.swapcase()
print(swapped_text)  # 输出: pYtHoN pRoGrAmMiNg
```

5. 去除空白字符

去除字符串中的空白符可以使用三种常见的方法：`strip()`、`lstrip()` 和 `rstrip()`。这些方法分别用于去除字符串两端或特定一端的空白符或其他字符。

`strip()` 方法用于去除字符串开头和结尾的所有空白符（包括空格、换行符、制表符等）。示例如下：

```
text = "   Python is great!   "
trimmed_text = text.strip()
print(trimmed_text)
# 输出: "Python is great!"
```

此方法不会影响字符串中间的空白符，只会去除两端的空白符。

`lstrip()` 方法用于去除字符串左侧的空白符，右侧保持不变：

```
text = "   Python is great!   "
left_trimmed_text = text.lstrip()
print(left_trimmed_text)
# 输出: "Python is great!   "
```

`rstrip()` 方法去除字符串右侧的空白符，左侧保持不变：

```
text = "   Python is great!   "
right_trimmed_text = text.rstrip()
print(right_trimmed_text)
# 输出: "   Python is great!"
```

字符串的 `count()` 方法用于计算指定子字符串在目标字符串中出现的次数。该方法非常适合用于文本处理和字符串分析任务，尤其是在需要统计某个字符或子字符串出现频率时。

```
string.count(substring, start=..., end=...)
```

- `substring`：必选参数，表示需要计数的子字符串。
- `start`（可选）：指定搜索的起始索引，默认为字符串的开头。
- `end`（可选）：指定搜索的结束索引，默认为字符串的末尾。

该方法返回一个整数，表示子字符串在指定范围内出现的次数。如果未找到子字符串，则返回 0。

`start` 和 `end` 参数含义同字符串切片。

示例 1：计数字符串中某字符的出现次数

```
message = 'python is popular programming language'
print(message.count('p'))  # 输出: 4
```

在上述代码中，`'p'` 在字符串中总共出现了 4 次。

示例 2：使用 `start` 和 `end` 参数

```
string = "Python is awesome, isn't it?"
substring = "i"
count = string.count(substring, 8, 25)
print("The count is:", count)  # 输出: 1
count = string.count(substring, 8, 26)
print("The count is:", count)  # 输出: 2
```

在这个示例中，计数从索引 8 开始，到索引 25 结束，因此只找到 1 次 `'i'` 字符。

字符串有多个以 `is` 开头的函数，这些函数用于对字符串内容进行各种类型的验证，返回布尔值（`True` 或 `False`）。如下表

表 1: 常见的字符串内容类型验证函数

函数	含义	示例代码	输出
`isalnum()`	判断字符串是否只包含字母和数字	`"Hello123".isalnum()`	`True`
`isalpha()`	判断字符串是否只包含字母	`"Hello".isalpha()`	`True`
`isdigit()`	判断字符串是否只包含数字	`"12345".isdigit()`	`True`
`isdecimal()`	判断字符串是否只包含十进制字符	`"12345".isdecimal()`	`True`
`islower()`	判断字符串是否全为小写字母	`"hello".islower()`	`True`
`isupper()`	判断字符串是否全为大写字母	`"HELLO".isupper()`	`True`
`istitle()`	判断字符串是否每个单词首字母大写）	`"Hello World".istitle()`	`True`
`isspace()`	判断字符串是否只包含空白字符	`" ".isspace()`	`True`

字符串还有许多实用的常用方法，参见讲义中的表 5.1，可以结合 `help` 函数自行学习其他常用的字符串方法的用法。

字符串格式化是一项重要的技能，特别是在处理动态文本输出时，如生成报告、用户提示或数据展示。Python 提供了多种格式化字符串的方法，包括旧式的百分号格式化 (%)，`str.format()` 方法，以及较新的 F 字符串格式化 (`f-strings`)。

1. 百分号格式化

这是 Python 最早的字符串格式化方式，通过使用% 符号替换占位符。例如：

```
name = "Alice"
age = 30
print("Hello, my name is %s and I am %d years old." % (name,
↪  age))
```

上例中，%s 表示字符串占位符，%d 表示整数占位符。该方法虽然简洁，但可读性和灵活性较低，已逐渐被 `str.format()` 和 `f-strings` 所取代。

2. `str.format()` 方法

`str.format()` 引入了更加灵活的字符串格式化方式。使用大括号 `{}` 作为占位符，支持位置参数和关键字参数。例如：

```
name = "Bob"
score = 95.5
message = "Student: {} | Score: {:.2f}".format(name, score)
print(message)
```

`{}` 占位符被 `name` 替换，而 `:.2f` 将 `score` 格式化为保留两位小数的浮点数。

3. F 字符串格式化 (`f-strings`)

Python 3.6 引入了 F 字符串格式化，这是目前推荐的格式化方式。它允许在字符串中直接嵌入变量和表达式，使代码更加简洁明了。例如：

```
name = "Eve"
gpa = 3.8
message = f"Student: {name} | GPA: {gpa:.2f}"
print(message)
```

变量 `name` 和 `gpa` 直接嵌入到字符串中，并且可以通过 `{gpa:.2f}` 将 `gpa` 格式化为两位小数的浮点数。F 字符串不仅支持变量插值，还能嵌入复杂的表达式。

str.format() 方法的位置参数和关键字参数

`str.format()` 方法可以通过位置参数和关键字参数来进行字符串格式化，灵活控制字符串的内容替换。

1. 位置参数

使用位置参数时，根据参数在 `format()` 方法中的顺序将值插入到字符串的占位符中，参数的顺序由大括号中的数字索引来决定。例如：

```
message = "Hello, {0}. You are {1} years old.".format("Alice",
↪  25)
print(message)
```

在这个例子中，`{0}` 和 `{1}` 分别表示 `"Alice"` 和 `25` 两个位置参数。

2. 关键字参数

关键字参数允许通过名称引用参数值，这样使代码更加清晰。例如：

```
message = "Hello, {name}. You are {age} years
↪  old.".format(name="Bob", age=30)
print(message)
```

通过使用关键字参数 `name` 和 `age`，可以指定各自的值，使得格式化更加直观。

可以混合使用位置参数和关键字参数，但要注意，位置参数**必须**在关键字参数之前。例如：

```
message = "Hello, {0}. Your balance is
↪  {balance}.".format("David", balance=230.23)
print(message)
```

Python 的 `string` 模块提供了一系列用于处理字符串的常量和函数。该模块包含常用的字符集合，如字母、数字、标点符号等，简化了字符串操作。此外，`string` 模块还提供了诸如 `capwords()`、`translate()` 等实用函数，能够实现字符转换、格式化等功能，特别适合在数据处理和文本清理中使用。

`string` 模块中，常量提供了一些预定义的字符集合，用于简化字符串处理。以下是一些常用常量及其基本用法。

```python
import string
# 输出所有小写字母
print(" 小写字母:", string.ascii_lowercase)
# 输出所有大写字母
print(" 大写字母:", string.ascii_uppercase)
# 输出所有字母（包含大写和小写）
print(" 所有字母:", string.ascii_letters)
print(" 数字字符:", string.digits) # 输出数字字符
# 输出标点符号
print(" 标点符号:", string.punctuation)
```

`translate()` 函数用于基于一个翻译表（translation table）替换或移除字符串中的字符。该翻译表可以通过 `str.maketrans()` 方法创建，`translate()` 函数结合此表高效地执行字符映射操作。这个功能常用于数据清理或字符串替换等场景。

```
# 导入 string 模块
import string

# 创建一个翻译表，替换字符并移除特定字符
translation_table = str.maketrans("abc", "123", "d")

# 应用 translate 函数
text = "abcdef"
translated_text = text.translate(translation_table)

# 输出结果
print(" 原始文本:", text)
print(" 翻译后的文本:", translated_text)
```

在 Python 字符串处理过程中，特殊字符（special characters）是指那些不能直接表示或具有特殊含义的字符。为了在字符串中正确使用这些字符，通常需要使用转义字符（escape character）来避免语法错误或实现特定功能。转义字符以反斜杠（\）为前缀，后跟一个特定字符，来表示一个特殊的含义。

1. **换行符 \n**: 用于在字符串中插入一个换行。

```
print("Hello\nWorld")
```

2. **制表符 \t**: 用于插入一个水平制表符。

```
print("Hello\tWorld")
```

3. **单引号 `'` 和双引号 `"`**：当字符串使用单引号或双引号时，如果需要在字符串中包含相同类型的引号，需要使用转义字符。

```
print('It\'s a beautiful day')
print("He said, \"Python is awesome!\"")
```

4. **反斜杠 **：用于表示一个实际的反斜杠，因为单个反斜杠在 Python 中是转义字符。

```
print("This is a backslash: \\")
```

5. **回车符 \r 和退格符 \b**：\r 用于将光标移到行首，\b 则是退格符，删除前一个字符。

```
print("Hello\rWorld")  # 输出为 "Worldo"
print("Hello\b World")  # 输出为 "Hell World"
```

6. **原始字符串 `r` 或 `R`**：在需要保留反斜杠的情况下，可以通过在字符串前加 `r` 或 `R`，使反斜杠不被解释为转义字符。

```
print(r"C:\new_folder\test.txt")
```

```
# 使用转义字符打印带有引号的字符串
print("He said, \"Python is fun!\"")
# 输出: He said, "Python is fun!"

# 打印包含路径的字符串
print(r"C:\Users\username\Desktop")
# 输出: C:\Users\username\Desktop

# 使用换行符和制表符格式化输出
print("Name:\tJohn\nAge:\t25")
# 输出:
# Name:    John
# Age:     25
```

未完待续