**Python Regular Expression Tutorial**

Discover Python regular expressions: find basic and complex patterns, repetitions, or to do (non-)greedy matching, work with the re library and much more!

Regular expressions are used to identify whether a pattern exists in a given sequence of characters (string) or not. They help in manipulating textual data, which is often a pre-requisite for data science projects that involve text mining. You must have come across some application of regular expressions: they are used at the server side to validate the format of email addresses or password during registration, used for parsing text data files to find, replace or delete certain string, etc.

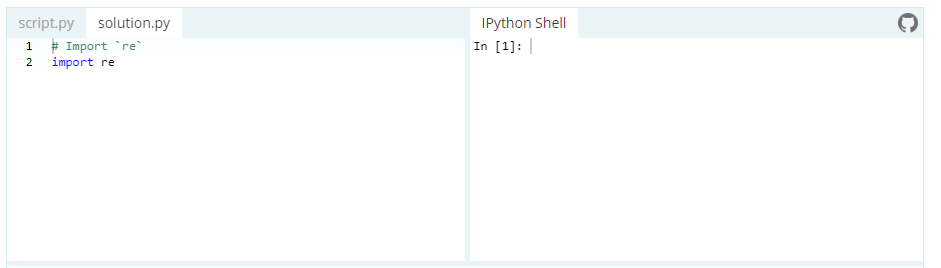
You see, regular expressions are extremely powerful and in this tutorial, you will learn to use them in Python. You will cover the following topics:

* [Regular Expressions in Python](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#REinPython)
* [Basic Characters: Ordinary Characters](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#BasicChar)
* [Wild Card Characters: Special Characters](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#WildChar)
* [Repetitions](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#repetitions)
* [Groups and Grouping using Regular Expressions](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#grouping)
* [Greedy vs Non-Greedy Matching](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#greedy)
* [re Python Library](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#modulecontent)
  + [search() versus match()](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#SearchVSMatch)

Check out DataCamp's [Natural Language Processing Fundamentals](https://www.datacamp.com/courses/natural-language-processing-fundamentals-in-python) course. This course dives deeper into using regular expressions in the context of solving common NLP problems. You will build a supervised learning classifier to identify "fake news". Be sure to try it out, the first chapter is free!

**Regular Expressions in Python**

In Python, regular expressions are supported by the [re](https://docs.python.org/3/library/re.html) module. That means that if you want to start using them in your Python scripts, you have to import this module with the help of import:

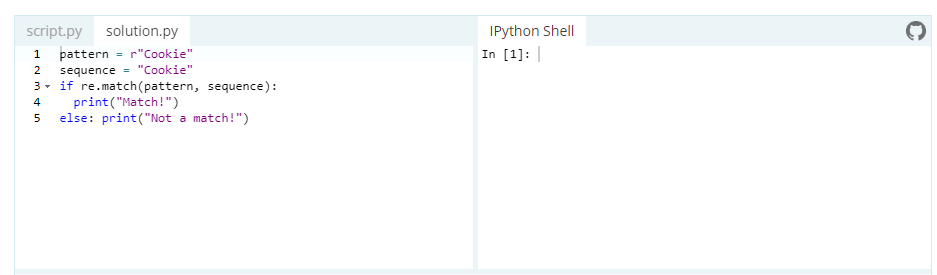


**Basic Patterns: Ordinary Characters**

You can easily tackle many basic patterns in Python using the ordinary characters. Ordinary characters are the simplest regular expressions. They match themselves exactly and do not have a special meaning in their regular expression syntax.

Examples are 'A', 'a', 'X', '5'.

Ordinary characters can be used to perform simple exact matches:



The match() function returns a match object if the text matches the pattern. Otherwise it returns None. The re module also contains several other [functions](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#modulecontent) and you will learn some of them later on in the tutorial.

For now, though, let's focus on ordinary characters! Do you notice the r at the start of the pattern Cookie?

This is called a *raw string literal*. It changes how the string literal is interpreted. Such literals are stored as they appear.

For example, \ is just a backslash when prefixed with a r rather than being interpreted as an escape sequence. You will see what this means with special characters. Sometimes, the syntax involves backslash-escaped characters and to prevent these characters from being interpreted as escape sequences, you use the raw r prefix. You don't actually need it for this example, however it is a good practice to use it for consistency.

**Wild Card Characters: Special Characters**

Special characters are characters which do not match themselves as seen but actually have a special meaning when used in a regular expression.

The most widely used special characters are:

* . - A period. Matches any single character except newline character.

re.search(r'Co.k.e', 'Cookie').group()

'Cookie'

The group() function returns the string matched by the re. You will see this function in more detail [later](https://www.datacamp.com/community/tutorials/python-regular-expression-tutorial#grouping).

* \w - Lowercase w. Matches any single letter, digit or underscore.

re.search(r'Co\wk\we', 'Cookie').group()

'Cookie'

* \W - Uppercase w. Matches any character not part of \w (lowercase w).

re.search(r'C\Wke', 'C@ke').group()

'C@ke'

* \s - Lowercase s. Matches a single whitespace character like: space, newline, tab, return.

re.search(r'Eat\scake', 'Eat cake').group()

'Eat cake'

* \S - Uppercase s. Matches any character not part of \s (lowercase s).

re.search(r'Cook\Se', 'Cookie').group()

'Cookie'

* \t - Lowercase t. Matches tab.

re.search(r'Eat\tcake', 'Eat cake').group()

'Eat\tcake'

* \n - Lowercase n. Matches newline.
* \r - Lowercase r. Matches return.
* \d - Lowercase d. Matches decimal digit 0-9.

re.search(r'c\d\dkie', 'c00kie').group()

'c00kie'

* ^ - Caret. Matches a pattern at the start of the string.

re.search(r'^Eat', 'Eat cake').group()

'Eat'

* $ - Matches a pattern at the end of string.

re.search(r'cake$', 'Eat cake').group()

'cake'

* [abc] - Matches a or b or c.
* [a-zA-Z0-9] - Matches any letter from (a to z) or (A to Z) or (0 to 9). Characters that are not within a range can be matched by complementing the set. If the first character of the set is ^, all the characters that are not in the set will be matched.

re.search(r'Number: [0-6]', 'Number: 5').group()

'Number: 5'

# Matches any character except 5

re.search(r'Number: [^5]', 'Number: 0').group()

'Number: 0'

* \A - Uppercase a. Matches only at the start of the string. Works across multiple lines as well.

re.search(r'\A[A-E]ookie', 'Cookie').group()

'Cookie'

* \b - Lowercase b. Matches only the beginning or end of the word.

re.search(r'\b[A-E]ookie', 'Cookie').group()

'Cookie'

* \ - Backslash. If the character following the backslash is a recognized escape character, then the special meaning of the term is taken. For example, \n is considered as newline. However, if the character following the \ is not a recognized escape character, then the \ is treated like any other character and passed through.

Let's look at a couple of examples:

# This checks for '\' in the string instead of '\t' due to the '\' used

re.search(r'Back\\stail', 'Back\stail').group()

'Back\\stail'

# This treats '\s' as an escape character because it lacks '\' at the start of '\s'

re.search(r'Back\stail', 'Back tail').group()

'Back lash'

**Repetitions**

It becomes quite tedious if you are looking to find long patterns in a sequence. Fortunately, the re module handles repetitions using the following special characters:

* + - Checks for one or more characters to its left.

re.search(r'Co+kie', 'Cooookie').group()

'Cooookie'

* \* - Checks for zero or more characters to its left.

# Checks for any occurrence of a or o or both in the given sequence

re.search(r'Ca\*o\*kie', 'Caokie').group()

'Caokie'

* ? - Checks for exactly zero or one character to its left.

# Checks for exactly zero or one occurrence of a or o or both in the given sequence

re.search(r'Colou?r', 'Color').group()

'Color'

But what if you want to check for exact number of sequence repetition?

For example, checking the validity of a phone number in an application. re module handles this very gracefully as well using the following regular expressions:

{x} - Repeat exactly x number of times.

{x,} - Repeat at least x times or more.

{x, y} - Repeat at least x times but no more than y times.

re.search(r'\d{9,10}', '0987654321').group()

'0987654321'

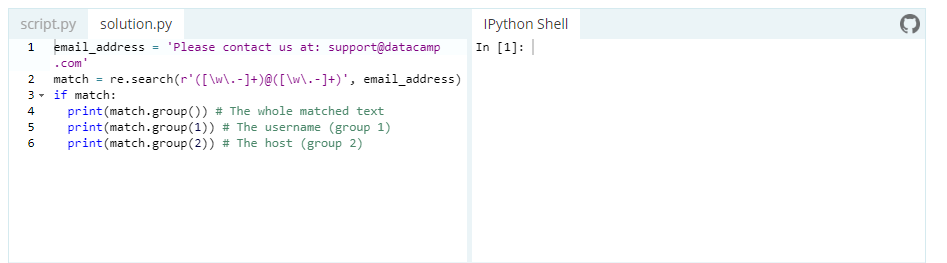
The + and \* qualifiers are said to be greedy.

**Groups and Grouping using Regular Expressions**

Suppose that, when you're validating email addresses and want to check the user name and host separately.

This is when the group feature of regular expression comes in handy. It allows you to pick up parts of the matching text.

Parts of a regular expression pattern bounded by parenthesis() are called groups. The parenthesis does not change what the expression matches, but rather forms groups within the matched sequence. You have been using the group() function all along in this tutorial's examples. The plain match.group() without any argument is still the whole matched text as usual.



**Greedy vs Non-Greedy Matching**

When a special character matches as much of the search sequence (string) as possible, it is said to be a "Greedy Match". It is the normal behavior of a regular expression but sometimes this behavior is not desired:

pattern = "cookie"

sequence = "Cake and cookie"

heading = r'<h1>TITLE</h1>'

re.match(r'<.\*>', heading).group()

'<h1>TITLE</h1>'

The pattern <.\*> matched the whole string, right up to the second occurrence of >.

However, if you only wanted to match the first <h1> tag, you could have used the greedy qualifier \*? that matches as little text as possible.

Adding ? after the qualifier makes it perform the match in a non-greedy or minimal fashion; That is, as few characters as possible will be matched. When you run <.\*>, you will only get a match with <h1>.

heading = r'<h1>TITLE</h1>'

re.match(r'<.\*?>', heading).group()

'<h1>'

**re Python Library**

The re library in Python provides several functions that makes it a skill worth mastering. You have already seen some of them, such as the re.search(), re.match(). Let's check out some useful functions in detail:

* search(pattern, string, flags=0)

With this function, you scan through the given string/sequence looking for the first location where the regular expression produces a match. It returns a corresponding match object if found, else returns None if no position in the string matches the pattern. Note that None is different from finding a zero-length match at some point in the string.

pattern = "cookie"

sequence = "Cake and cookie"

re.search(pattern, sequence).group()

'cookie'

* match(pattern, string, flags=0)

Returns a corresponding match object if zero or more characters at the beginning of string match the pattern. Else it returns None, if the string does not match the given pattern.

pattern = "C"

sequence1 = "IceCream"

# No match since "C" is not at the start of "IceCream"

re.match(pattern, sequence1)

sequence2 = "Cake"

re.match(pattern,sequence2).group()

'C'

**search() versus match()**

The match() function checks for a match only at the beginning of the string (by default) whereas the search() function checks for a match anywhere in the string.

* findall(pattern, string, flags=0)

Finds all the possible matches in the entire sequence and returns them as a list of strings. Each returned string represents one match.

email\_address = "Please contact us at: support@datacamp.com, xyz@datacamp.com"

#'addresses' is a list that stores all the possible match

addresses = re.findall(r'[\w\.-]+@[\w\.-]+', email\_address)

for address in addresses:

print(address)

support@datacamp.com

xyz@datacamp.com

* sub(pattern, repl, string, count=0, flags=0)

This is the substitute function. It returns the string obtained by replacing or substituting the leftmost non-overlapping occurrences of pattern in string by the replacement repl. If the pattern is not found then the string is returned unchanged.

email\_address = "Please contact us at: xyz@datacamp.com"

new\_email\_address = re.sub(r'([\w\.-]+)@([\w\.-]+)', r'support@datacamp.com', email\_address)

print(new\_email\_address)

Please contact us at: support@datacamp.com

* compile(pattern, flags=0)

Compiles a regular expression pattern into a [regular expression object](https://docs.python.org/3/library/re.html#re-objects). When you need to use an expression several times in a single program, using the compile() function to save the resulting regular expression object for reuse is more efficient. This is because the compiled versions of the most recent patterns passed to compile() and the module-level matching functions are cached.

pattern = re.compile(r"cookie")

sequence = "Cake and cookie"

pattern.search(sequence).group()

'cookie'

# This is equivalent to:

re.search(pattern, sequence).group()

'cookie'

**Tip** : an expression's behavior can be modified by specifying a flags value. You can add flag as an extra argument to the various functions that you have seen in this tutorial. Some of the flags used are: IGNORECASE, DOTALL, MULTILINE, VERBOSE, etc.

**Case Study: Working with Regular Expressions**

Now that you have seen how regular expressions work in Python by studying some examples, it's time to get your hands dirty! In this case study, you'll put your knowledge to work.

import re

import requests

the\_idiot\_url = 'https://www.gutenberg.org/files/2638/2638-0.txt'

def get\_book(url):

# Sends a http request to get the text from project Gutenberg

raw = requests.get(url).text

# Discards the metadata from the beginning of the book

start = re.search(r"\\*\\*\\* START OF THIS PROJECT GUTENBERG EBOOK .\* \\*\\*\\*",raw ).end()

# Discards the metadata from the end of the book

stop = re.search(r"II", raw).start()

# Keeps the relevant text

text = raw[start:stop]

return text

def preprocess(sentence):

return re.sub('[^A-Za-z0-9.]+' , ' ', sentence).lower()

book = get\_book(the\_idiot\_url)

processed\_book = preprocess(book)

print(processed\_book)

Find the number of the pronoun "the" in the corpus. Hint: use the len() function.

len(re.findall(r'the', processed\_book))

302

Try to convert every single stand-alone instance of 'i' to 'I' in the corpus. Make sure not to change the 'i' occuring in a word:

processed\_book = re.sub(r'\si\s', " I ", processed\_book)

print(processed\_book)

Find the number of times anyone was quoted ("") in the corpus.

len(re.findall(r'\”', book))

96

What are the words connected by '--' in the corpus?

re.findall(r'[a-zA-Z0-9]\*--[a-zA-Z0-9]\*', book)

['ironical--it',

'malicious--smile',

'fur--or',

'astrachan--overcoat',

'it--the',

'Italy--was',

'malady--a',

'money--and',

'little--to',

'No--Mr',

'is--where',

'I--I',

'I--',

'--though',

'crime--we',

'or--judge',

'gaiters--still',

'--if',

'through--well',

'say--through',

'however--and',

'Epanchin--oh',

'too--at',

'was--and',

'Andreevitch--that',

'everyone--that',

'reduce--or',

'raise--to',

'listen--and',

'history--but',

'individual--one',

'yes--I',

'but--',

't--not',

'me--then',

'perhaps--',

'Yes--those',

'me--is',

'servility--if',

'Rogojin--hereditary',

'citizen--who',

'least--goodness',

'memory--but',

'latter--since',

'Rogojin--hung',

'him--I',

'anything--she',

'old--and',

'you--scarecrow',

'certainly--certainly',

'father--I',

'Barashkoff--I',

'see--and',

'everything--Lebedeff',

'about--he',

'now--I',

'Lihachof--',

'Zaleshoff--looking',

'old--fifty',

'so--and',

'this--do',

'day--not',

'that--',

'do--by',

'know--my',

'illness--I',

'well--here',

'fellow--you']