

Python re(gex)?

a magical tool for text processing

Give me all leaves



Sundeeep Agarwal

Table of contents

Preface	5
Prerequisites	5
Conventions	5
Acknowledgements	5
Feedback and Errata	6
Author info	6
License	6
Book version	6
Why is it needed?	7
re introduction	8
re module documentation	8
re.search	8
re.search in conditional expressions	9
re.sub	9
Compiling regular expressions	10
bytes	11
Cheatsheet and Summary	12
Exercises	12
Anchors	15
String anchors	15
re.fullmatch	16
Line anchors	17
Word anchors	18
Cheatsheet and Summary	20
Exercises	20
Alternation and Grouping	23
Alternation	23
Grouping	24
Precedence rules	25
Cheatsheet and Summary	27
Exercises	27
Escaping metacharacters	29
Escaping with \	29
re.escape	29
Escape sequences	30
Cheatsheet and Summary	31
Exercises	31
Dot metacharacter and Quantifiers	34
Dot metacharacter	34
re.split	34
Greedy quantifiers	35
Conditional AND	37
What does greedy mean?	38

Non-greedy quantifiers	39
Cheatsheet and Summary	39
Exercises	40
Interlude: Tools for debugging and visualization	43
regex101	43
debuggex	43
regexcrossword	44
Summary	44
Working with matched portions	45
re.Match object	45
Assignment expressions	47
Using functions in replacement section	47
Using dict in replacement section	48
re.findall	48
re.finditer	49
re.split with capture groups	50
re.subn	51
Cheatsheet and Summary	52
Exercises	53
Character class	57
Custom character sets	57
Range of characters	57
Negating character sets	58
Matching metacharacters literally	58
Escape sequence sets	59
Numeric ranges	60
Cheatsheet and Summary	61
Exercises	61
Groupings and backreferences	65
Backreference	65
Non-capturing groups	67
Named capture groups	68
Conditional groups	69
Match.expand	70
Cheatsheet and Summary	70
Exercises	71
Interlude: Common tasks	75
CommonRegex	75
Summary	75
Lookarounds	76
Conditional expressions	76
Negative lookarounds	76
Positive lookarounds	78
Capture groups inside positive lookarounds	78
Conditional AND with lookarounds	79

Variable length lookbehind	79
Negated groups	80
Cheatsheet and Summary	81
Exercises	82
Flags	86
re.IGNORECASE	86
re.DOTALL	86
re.MULTILINE	87
re.VERBOSE	87
Inline comments	88
Inline flags	88
Cheatsheet and Summary	89
Exercises	89
Unicode	92
re.ASCII	92
Codepoints and Unicode escapes	93
\N escape sequence	93
Cheatsheet and Summary	93
Exercises	94
regex module	95
Possessive quantifiers	95
Subexpression calls	96
Positive lookbehind with \K	96
Variable length lookbehind	97
\G anchor	98
Recursive matching	98
Named character sets	100
Set operations	100
Unicode character sets	101
Skipping matches	101
\m and \M word anchors	102
Overlapped matches	102
regex.REVERSE flag	103
\X vs dot metacharacter	103
Cheatsheet and Summary	103
Exercises	104
Gotchas	109
Escape sequences	109
Line anchors with \n as last character	109
Zero length matches	110
Capture group with quantifiers	110
Converting re to regex module	110
Optional arguments syntax	111
Summary	111
Further Reading	112

Preface

Scripting and automation tasks often need to extract particular portions of text from input data or modify them from one format to another. This book will help you learn Regular Expressions, a mini-programming language for all sorts of text processing needs.

The book heavily leans on examples to present features of regular expressions one by one. It is recommended that you manually type each example and experiment with them. Understanding both the nature of sample input string and the output produced is essential. As an analogy, consider learning to drive a bike or a car — no matter how much you read about them or listen to explanations, you need to practice a lot and infer your own conclusions. Should you feel that copy-paste is ideal for you, [code snippets are available chapter wise on GitHub](#).

Prerequisites

Prior experience working with Python, should know concepts like string formats, string methods, list comprehension and so on.

If you have prior experience with a programming language, but new to Python, see my list of [Python curated resources](#) before starting this book.

Conventions

- The examples presented here have been tested with **Python version 3.8.3** and includes features not available in earlier versions.
- Code snippets shown are copy pasted from Python REPL shell and modified for presentation purposes. Some commands are preceded by comments to provide context and explanations. Blank lines have been added to improve readability. Error messages are shortened. `import` statements are skipped after initial use. And so on.
- Unless otherwise noted, all examples and explanations are meant for **ASCII** characters.
- External links are provided for further reading throughout the book. Not necessary to immediately visit them. They have been chosen with care and would help, especially during re-reads.
- The [py_regular_expressions repo](#) has all the code snippets and files used in examples and exercises and other details related to the book. If you are not familiar with `git` command, click the **Clone** button on the webpage to get the files.

Acknowledgements

- [Python documentation](#) — manuals and tutorials
- [/r/learnpython/](#) and [/r/regex/](#) — helpful forums for beginners and experienced programmers alike
- [stackoverflow](#) — for getting answers to pertinent questions on Python and regular expressions
- [tex.stackexchange](#) — for help on `pandoc` and `tex` related questions
- Cover image: [draw.io](#), [tree icon](#) by [Gopi Doraisamy](#) under [Creative Commons Attribution 3.0 Unported](#) and [wand icon](#) by [roundicons.com](#)

- [Warning](#) and [Info](#) icons by [Amada44](#) under public domain
- [David Cortesi](#) for helpful feedback on both the technical content and grammar issues
- **Kye** for spotting a typo
- **Hugh**'s email exchanges helped me significantly to improve the presentation of concepts and exercises
- [Christopher Patti](#) for reviewing the book, providing feedback and brightening the day with kind words

Special thanks to Al Sweigart, for introducing me to Python with his awesome [automatethe-boringstuff](#) book and video course.

Feedback and Errata

I would highly appreciate if you'd let me know how you felt about this book, it would help to improve this book as well as my future attempts. Also, please do let me know if you spot any error or typo.

Issue Manager: https://github.com/learnbyexample/py_regular_expressions/issues

Goodreads: <https://www.goodreads.com/book/show/47142552-python-re-gex>

E-mail: learnbyexample.net@gmail.com

Twitter: https://twitter.com/learn_byexample

Author info

Sundeeep Agarwal is a freelance trainer, author and mentor. His previous experience includes working as a Design Engineer at Analog Devices for more than 5 years. You can find his other works, primarily focused on Linux command line, text processing, scripting languages and curated lists, at <https://github.com/learnbyexample>. He has also been a technical reviewer for [Command Line Fundamentals](#) book and video course published by Packt.

List of books: <https://learnbyexample.github.io/books/>

License

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#)

Code snippets are available under [MIT License](#)

Resources mentioned in Acknowledgements section above are available under original licenses.

Book version

3.0

See [Version_changes.md](#) to track changes across book versions.

Why is it needed?

Regular Expressions is a versatile tool for text processing. You'll find them included as part of standard library of most programming languages that are used for scripting purposes. If not, you can usually find a third-party library. Syntax and features of regular expressions vary from language to language. Python's syntax is similar to that of Perl language, but there are significant feature differences.

The `str` class comes loaded with variety of methods to deal with text. So, what's so special about regular expressions and why would you need it? For learning and understanding purposes, one can view regular expressions as a mini programming language in itself, specialized for text processing. Parts of a regular expression can be saved for future use, analogous to variables and functions. There are ways to perform AND, OR, NOT conditionals. Operations similar to range function, string repetition operator and so on.

Here's some common use cases.

- Sanitizing a string to ensure that it satisfies a known set of rules. For example, to check if a given string matches password rules.
- Filtering or extracting portions on an abstract level like alphabets, numbers, punctuation and so on.
- Qualified string replacement. For example, at the start or the end of a string, only whole words, based on surrounding text, etc.

Further Reading

- [The true power of regular expressions](#) — it also includes a nice explanation of what *regular* means in this context
- [softwareengineering: Is it a must for every programmer to learn regular expressions?](#)
- [softwareengineering: When you should NOT use Regular Expressions?](#)
- [codinghorror: Now You Have Two Problems](#)
- [wikipedia: Regular expression](#) — this article includes discussion on regular expressions as a formal language as well as details on various implementations

re introduction

In this chapter, you'll get an introduction of `re` module that is part of Python's standard library. For some examples, the equivalent normal string method is shown for comparison. This chapter focuses on syntax, regular expression features will be covered next chapter onwards.

re module documentation

It is always a good idea to know where to find the documentation. The default offering for Python regular expressions is the `re` standard library module. Visit [docs.python: re](https://docs.python.org/3/library/re.html) for information on available methods, syntax, features, examples and more. Here's a quote:

A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression

re.search

Normally you'd use the `in` operator to test whether a string is part of another string or not. For regular expressions, use the `re.search` function whose argument list is shown below.

```
re.search(pattern, string, flags=0)
```

The first argument is the RE pattern you want to test against the input string, which is the second argument. `flags` is optional, it helps to change the default behavior of RE patterns.

As a good practice, always use **raw strings** to construct the RE pattern. This will become clearer in later chapters. Here's some examples.

```
>>> sentence = 'This is a sample string'

# check if 'sentence' contains the given search string
>>> 'is' in sentence
True
>>> 'xyz' in sentence
False

# need to load the re module before use
>>> import re

# check if 'sentence' contains the pattern described by RE argument
>>> bool(re.search(r'is', sentence))
True
>>> bool(re.search(r'xyz', sentence))
False
```

Before using the `re` module, you need to `import` it. Further example snippets will assume that the module is already loaded. The return value of `re.search` function is a `re.Match`

object when a match is found and `None` otherwise (note that I treat `re` as a word, not as `r` and `e` separately, hence the use of *a* instead of *an*). More details about the `re.Match` object will be discussed in [Working with matched portions](#) chapter. For presentation purposes, the examples will use `bool` function to show `True` or `False` depending on whether the RE pattern matched or not.

Here's an example with `flags` optional argument. It will be discussed in detail in [Flags](#) chapter.

```
>>> sentence = 'This is a sample string'

>>> bool(re.search(r'this', sentence))
False

# re.IGNORECASE (or re.I) is a flag to enable case insensitive matching
>>> bool(re.search(r'this', sentence, flags=re.I))
True
```

re.search in conditional expressions

As Python evaluates `None` as `False` in boolean context, `re.search` can be used directly in conditional expressions. See also [docs.python: Truth Value Testing](#).

```
>>> sentence = 'This is a sample string'
>>> if re.search(r'ring', sentence):
...     print('mission success')
...
mission success

>>> if not re.search(r'xyz', sentence):
...     print('mission failed')
...
mission failed
```

Here's some generator expression examples.

```
>>> words = ['cat', 'attempt', 'tattle']

>>> [w for w in words if re.search(r'tt', w)]
['attempt', 'tattle']
>>> all(re.search(r'at', w) for w in words)
True
>>> any(re.search(r'stat', w) for w in words)
False
```

re.sub

For normal search and replace, you'd use the `str.replace` method. For regular expressions, use the `re.sub` function, whose argument list is shown below.

```
re.sub(pattern, repl, string, count=0, flags=0)
```

The first argument is the RE pattern to match against the input string, which is the third argument. The second argument specifies the string which will replace the portions matched by the RE pattern. `count` and `flags` are optional arguments.

```
>>> greeting = 'Have a nice weekend'

# replace all occurrences of 'e' with 'E'
# same as: greeting.replace('e', 'E')
>>> re.sub(r'e', 'E', greeting)
'HavE a nicE wEEkEnd'

# replace first two occurrences of 'e' with 'E'
# same as: greeting.replace('e', 'E', 2)
>>> re.sub(r'e', 'E', greeting, count=2)
'HavE a nicE weekend'
```



A common mistake, not specific to `re.sub`, is forgetting that strings are immutable in Python.

```
>>> word = 'cater'
# this will return a string object, won't modify 'word' variable
>>> re.sub(r'cat', 'wag', word)
'wager'
>>> word
'cater'

# need to explicitly assign the result if 'word' has to be changed
>>> word = re.sub(r'cat', 'wag', word)
>>> word
'wager'
```

Compiling regular expressions

Regular expressions can be compiled using `re.compile` function, which gives back a `re.Pattern` object.

```
re.compile(pattern, flags=0)
```

The top level `re` module functions are all available as methods for such objects. Compiling a regular expression is useful if the RE has to be used in multiple places or called upon multiple times inside a loop (speed benefit).



By default, Python maintains a small list of recently used RE, so the speed benefit doesn't apply for trivial use cases. See also [stackoverflow: Is it worth using re.compile?](#)

```
>>> pet = re.compile(r'dog')
>>> type(pet)
<class 're.Pattern'>

# note that 'search' is called upon 'pet' which is a 're.Pattern' object
# since 'pet' has the RE information, you only need to pass input string
>>> bool(pet.search('They bought a dog'))
True
>>> bool(pet.search('A cat crossed their path'))
False

# replace all occurrences of 'dog' with 'cat'
>>> pet.sub('cat', 'They bought a dog')
'They bought a cat'
```

Some of the methods available for compiled patterns also accept more arguments than those available for top level functions of the `re` module. For example, the `search` method on a compiled pattern has two optional arguments to specify **start** and **end** index positions. Similar to `range` function and slicing notation, the ending index has to be specified 1 greater than desired index.

```
Pattern.search(string[, pos[, endpos]])
```

Note that there's no `flags` option as that has to be specified with `re.compile`.

```
>>> sentence = 'This is a sample string'
>>> word = re.compile(r'is')

# search for 'is' starting from 5th character of 'sentence' variable
>>> bool(word.search(sentence, 4))
True

# search for 'is' starting from 7th character of 'sentence' variable
>>> bool(word.search(sentence, 6))
False

# search for 'is' between 3rd and 4th characters
>>> bool(word.search(sentence, 2, 4))
True
```

bytes

To work with `bytes` data type, the RE must be of `bytes` data as well. Similar to `str` RE, use **raw** format to construct a `bytes` RE.

```
>>> byte_data = b'This is a sample string'

# error message truncated for presentation purposes
>>> re.search(r'is', byte_data)
TypeError: cannot use a string pattern on a bytes-like object

# use rb'..' for constructing bytes pattern
>>> bool(re.search(rb'is', byte_data))
True
>>> bool(re.search(rb'xyz', byte_data))
False
```

Cheatsheet and Summary

Note	Description
docs.python: re	Python standard module for regular expressions
<code>re.search</code>	Check if given pattern is present anywhere in input string <code>re.search(pattern, string, flags=0)</code> Output is a <code>re.Match</code> object, usable in conditional expressions raw strings preferred to define RE
<code>re.sub</code>	Additionally, Python maintains a small cache of recent RE search and replace using RE <code>re.sub(pattern, repl, string, count=0, flags=0)</code>
<code>re.compile</code>	Compile a pattern for reuse, output is a <code>re.Pattern</code> object <code>re.compile(pattern, flags=0)</code>
<code>rb'pat'</code>	Use byte pattern for byte input
<code>re.IGNORECASE</code> or <code>re.I</code>	flag to ignore case while matching

This chapter introduced the `re` module, which is part of the standard library. Functions `re.search` and `re.sub` were discussed as well as how to compile RE using `re.compile` function. The RE pattern is usually defined using raw strings. For byte input, the pattern has to be of byte type too. Although the `re` module is good enough for most use cases, there are situations where you need to use the third party `regex` module. To avoid mixing up features, a separate chapter is dedicated for the [regex module](#) towards the end of the book.

The next section has exercises to test your understanding of the concepts introduced in this chapter. Please do solve them before moving on to the next chapter.

Exercises



Try to solve exercises in every chapter using only the features discussed until that chapter. Some of the exercises will be easier to solve with techniques presented in later chapters, but the aim of these exercises is to explore the features presented so far.

a) Check whether the given strings contain `0xB0` . Display a boolean result as shown below.

```
>>> line1 = 'start address: 0xA0, func1 address: 0xC0'
>>> line2 = 'end address: 0xFF, func2 address: 0xB0'

>>> bool(re.search(r'', line1))      ##### add your solution here
False
>>> bool(re.search(r'', line2))      ##### add your solution here
True
```

b) Replace all occurrences of `5` with `five` for the given string.

```
>>> ip = 'They ate 5 apples and 5 oranges'

>>> re.sub()      ##### add your solution here
'They ate five apples and five oranges'
```

c) Replace first occurrence of `5` with `five` for the given string.

```
>>> ip = 'They ate 5 apples and 5 oranges'

>>> re.sub()      ##### add your solution here
'They ate five apples and 5 oranges'
```

d) For the given list, filter all elements that do *not* contain `e` .

```
>>> items = ['goal', 'new', 'user', 'sit', 'eat', 'dinner']

>>> [w for w in items if not re.search()]      ##### add your solution here
['goal', 'sit']
```

e) Replace all occurrences of `note` irrespective of case with `X` .

```
>>> ip = 'This note should not be NoTeD'

>>> re.sub()      ##### add your solution here
'This X should not be XD'
```

f) Check if `at` is present in the given byte input data.

```
>>> ip = b'tiger imp goat'

>>> bool(re.search())      ##### add your solution here
True
```

g) For the given input string, display all lines not containing `start` irrespective of case.

```
>>> para = '''good start
... Start working on that
... project you always wanted
... stars are shining brightly
... hi there
... start and try to
... finish the book
... bye'''
```

```
>>> pat = re.compile()          ##### add your solution here
>>> for line in para.split('\n'):
...     if not pat.search(line):
...         print(line)
...
project you always wanted
stars are shining brightly
hi there
finish the book
bye
```

h) For the given list, filter all elements that contains either `a` or `w` .

```
>>> items = ['goal', 'new', 'user', 'sit', 'eat', 'dinner']

##### add your solution here
>>> [w for w in items if re.search() or re.search()]
['goal', 'new', 'eat']
```

i) For the given list, filter all elements that contains both `e` and `n` .

```
>>> items = ['goal', 'new', 'user', 'sit', 'eat', 'dinner']

##### add your solution here
>>> [w for w in items if re.search() and re.search()]
['new', 'dinner']
```

j) For the given string, replace `0xA0` with `0x7F` and `0xC0` with `0x1F` .

```
>>> ip = 'start address: 0xA0, func1 address: 0xC0'

##### add your solution here
'start address: 0x7F, func1 address: 0x1F'
```

Anchors

Now that you're familiar with RE syntax and couple of `re` module functions, the next step is to know about the special features of regular expressions. In this chapter, you'll be learning about qualifying a pattern. Instead of matching anywhere in the given input string, restrictions can be specified. For now, you'll see the ones that are already part of `re` module. In later chapters, you'll learn how to define your own rules for restriction.

These restrictions are made possible by assigning special meaning to certain characters and escape sequences. The characters with special meaning are known as **metacharacters** in regular expressions parlance. In case you need to match those characters literally, you need to escape them with a `\` character (discussed in [Escaping metacharacters](#) chapter).

String anchors

This restriction is about qualifying a RE to match only at the start or the end of an input string. These provide functionality similar to the `str` methods `startswith` and `endswith`. First up, the escape sequence `\A` which restricts the matching to the start of string.

```
# \A is placed as a prefix to the search term
>>> bool(re.search(r'\Acat', 'cater'))
True
>>> bool(re.search(r'\Acat', 'concatenation'))
False

>>> bool(re.search(r'\Ahi', 'hi hello\n\top spot'))
True
>>> bool(re.search(r'\Atop', 'hi hello\n\top spot'))
False
```

To restrict the matching to the end of string, `\Z` is used.

```
# \Z is placed as a suffix to the search term
>>> bool(re.search(r'are\Z', 'spare'))
True
>>> bool(re.search(r'are\Z', 'nearest'))
False

>>> words = ['surrender', 'unicorn', 'newer', 'door', 'empty', 'eel', 'pest']
>>> [w for w in words if re.search(r'er\Z', w)]
['surrender', 'newer']
>>> [w for w in words if re.search(r't\Z', w)]
['pest']
```

You can emulate string concatenation operations by using the anchors by themselves as a pattern.

```
# insert text at the start of a string
>>> re.sub(r'\A', 're', 'live')
'relive'
>>> re.sub(r'\A', 're', 'send')
```

```
'resend'

# appending text
>>> re.sub(r'\Z', 'er', 'cat')
'cater'
>>> re.sub(r'\Z', 'er', 'hack')
'hacker'
```



Use the optional start and end index arguments for `Pattern.search` method with caution. They are not equivalent to string slicing. For example, specifying a greater than 0 start index when using `\A` is always going to return `False`. This is because, as far as the `search` method is concerned, only the search space is narrowed and the anchor positions haven't changed. When slicing is used, you are creating an entirely new string object with new anchor positions.

```
>>> word_pat = re.compile(r'\Aat')

>>> bool(word_pat.search('cater', 1))
False
>>> bool(word_pat.search('cater'[1:]))
True
```

re.fullmatch

Combining both the start and end string anchors, you can restrict the matching to the whole string. Similar to comparing strings using the `==` operator.

```
>>> word_pat = re.compile(r'\Acat\Z')

>>> bool(word_pat.search('cat'))
True
>>> bool(word_pat.search('concatenation'))
False
```

You can also use `re.fullmatch` function to ensure the pattern matches only the whole input string and not just a part of the input. This may not seem useful with features introduced so far, but when you have a complex RE pattern with multiple alternatives, this function is quite handy. The argument list is same as the `re.search` function.

```
re.fullmatch(pattern, string, flags=0)
```

```
>>> word_pat = re.compile(r'cat', flags=re.I)

>>> bool(word_pat.fullmatch('Cat'))
True
>>> bool(word_pat.fullmatch('Scatter'))
False
```

Line anchors

A string input may contain single or multiple lines. The newline character `\n` is used as the line separator. There are two line anchors, `^` metacharacter for matching the start of line and `$` for matching the end of line. If there are no newline characters in the input string, these will behave same as `\A` and `\Z` respectively.

```
>>> pets = 'cat and dog'

>>> bool(re.search(r'^cat', pets))
True
>>> bool(re.search(r'^dog', pets))
False

>>> bool(re.search(r'dog$', pets))
True
>>> bool(re.search(r'^dog$', pets))
False
```

By default, the input string is considered as a single line, even if multiple newline characters are present. In such cases, the `$` metacharacter can match both the end of string and just before `\n` if it is the last character. However, `\Z` will always match the end of string, irrespective of the characters present.

```
>>> greeting = 'hi there\nhave a nice day\n'

>>> bool(re.search(r'day$', greeting))
True
>>> bool(re.search(r'day\n$', greeting))
True

>>> bool(re.search(r'day\Z', greeting))
False
>>> bool(re.search(r'day\n\Z', greeting))
True
```

To indicate that the input string should be treated as multiple lines, you need to enable the `re.MULTILINE` flag (or `re.M` short form).

```
# check if any line in the string starts with 'top'
>>> bool(re.search(r'^top', 'hi hello\ntop spot', flags=re.M))
True

# check if any line in the string ends with 'ar'
>>> bool(re.search(r'ar$', 'spare\npar\ndare', flags=re.M))
True

# filter all elements having lines ending with 'are'
>>> elements = ['spare\ntool', 'par\n', 'dare']
>>> [e for e in elements if re.search(r'are$', e, flags=re.M)]
['spare\ntool', 'dare']
```

```
# check if any complete line in the string is 'par'
>>> bool(re.search(r'^par$', 'spare\npar\ndare', flags=re.M))
True
```

Just like string anchors, you can use the line anchors by themselves as a pattern.

```
# note that there is no \n at the end of this input string
>>> ip_lines = 'catapults\nconcatenate\ncat'
>>> print(re.sub(r'^', '* ', ip_lines, flags=re.M))
* catapults
* concatenate
* cat

>>> print(re.sub(r'$', '.', ip_lines, flags=re.M))
catapults.
concatenate.
cat.
```



If you are dealing with Windows OS based text files, you'll have to convert `\r\n` line endings to `\n` first. Which is easily handled by many of the Python functions and methods. For example, you can specify which line ending to use for `open` function, the `split` string method handles all whitespaces by default and so on. Or, you can handle `\r` as optional character with quantifiers (see [Dot metacharacter and Quantifiers](#) chapter).

Word anchors

The third type of restriction is word anchors. Alphabets (irrespective of case), digits and the underscore character qualify as word characters. You might wonder why there are digits and underscores as well, why not only alphabets? This comes from variable and function naming conventions — typically alphabets, digits and underscores are allowed. So, the definition is more oriented to programming languages than natural ones.

The escape sequence `\b` denotes a word boundary. This works for both the start of word and end of word anchoring. Start of word means either the character prior to the word is a non-word character or there is no character (start of string). Similarly, end of word means the character after the word is a non-word character or no character (end of string). This implies that you cannot have word boundary `\b` without a word character.

```
>>> words = 'par spar apparent spare part'

# replace 'par' irrespective of where it occurs
>>> re.sub(r'par', 'X', words)
'X sX apXent sXe Xt'

# replace 'par' only at start of word
>>> re.sub(r'\bpar', 'X', words)
'X spar apparent spare Xt'

# replace 'par' only at end of word
```

```
>>> re.sub(r'par\b', 'X', words)
'X sX apparent spare part'
# replace 'par' only if it is not part of another word
>>> re.sub(r'\bpar\b', 'X', words)
'X spar apparent spare part'
```

You can get lot more creative with using word boundary as a pattern by itself:

```
# space separated words to double quoted csv
# note the use of 'replace' string method for normal string replacement
# 'translate' method can also be used
>>> words = 'par spar apparent spare part'
>>> print(re.sub(r'\b', '"', words).replace(' ', ','))
"par","spar","apparent","spare","part"

>>> re.sub(r'\b', ' ', '-----hello-----')
'----- hello -----'

# make a programming statement more readable
# shown for illustration purpose only, won't work for all cases
>>> re.sub(r'\b', ' ', 'foo_baz=num1+35*42/num2')
' foo_baz = num1 + 35 * 42 / num2 '
# excess space at start/end of string can be stripped off
# later you'll learn how to add a qualifier so that strip is not needed
>>> re.sub(r'\b', ' ', 'foo_baz=num1+35*42/num2').strip()
'foo_baz = num1 + 35 * 42 / num2'
```

The word boundary has an opposite anchor too. `\B` matches wherever `\b` doesn't match. This duality will be seen with some other escape sequences too. Negative logic is handy in many text processing situations. But use it with care, you might end up matching things you didn't intend!

```
>>> words = 'par spar apparent spare part'

# replace 'par' if it is not start of word
>>> re.sub(r'\Bpar', 'X', words)
'par sX apXent sXe part'
# replace 'par' at end of word but not whole word 'par'
>>> re.sub(r'\Bpar\b', 'X', words)
'par sX apparent spare part'
# replace 'par' if it is not end of word
>>> re.sub(r'par\b', 'X', words)
'par spar apXent sXe Xt'
# replace 'par' if it is surrounded by word characters
>>> re.sub(r'\Bpar\b', 'X', words)
'par spar apXent sXe part'
```

Here's some standalone pattern usage to compare and contrast the two word anchors.

```
>>> re.sub(r'\b', ': ', 'copper')
':copper:'
```

```
>>> re.sub(r'\B', ':', 'copper')
'c:o:p:p:e:r'

>>> re.sub(r'\b', ' ', '-----hello-----')
'----- hello -----'

>>> re.sub(r'\B', ' ', '-----hello-----')
' - - - - -h e l l o- - - - - '
```

Cheatsheet and Summary

Note	Description
<code>\A</code>	restricts the match to the start of string
<code>\Z</code>	restricts the match to the end of string
<code>re.fullmatch</code>	ensures pattern matches the entire input string <code>re.fullmatch(pattern, string, flags=0)</code>
<code>\n</code>	line separator, dos-style files need special attention
metacharacter	characters with special meaning in RE
<code>^</code>	restricts the match to the start of line
<code>\$</code>	restricts the match to the end of line
<code>re.MULTILINE</code> or <code>re.M</code>	flag to treat input as multiline string
<code>\b</code>	restricts the match to the start/end of words word characters: alphabets, digits, underscore
<code>\B</code>	matches wherever <code>\b</code> doesn't match

In this chapter, you've begun to see building blocks of regular expressions and how they can be used in interesting ways. But at the same time, regular expression is but another tool in the land of text processing. Often, you'd get simpler solution by combining regular expressions with other string methods and generator expressions. Practice, experience and imagination would help you construct creative solutions. In coming chapters, you'll see more applications of anchors. The `regex` module also supports [\G anchor](#) which is best understood in combination with other regular expression features.

Exercises

a) Check if the given strings start with `be` .

```
>>> line1 = 'be nice'
>>> line2 = '"best!'"
>>> line3 = 'better?'
>>> line4 = 'oh no\nbear spotted'

>>> pat = re.compile()          ##### add your solution here

>>> bool(pat.search(line1))
True
>>> bool(pat.search(line2))
```

```
False
>>> bool(pat.search(line3))
True
>>> bool(pat.search(line4))
False
```

b) For the given input string, change only whole word `red` to `brown`

```
>>> words = 'bred red spread credible'

>>> re.sub() ##### add your solution here
'bred brown spread credible'
```

c) For the given input list, filter all elements that contains `42` surrounded by word characters.

```
>>> words = ['hi42bye', 'nice1423', 'bad42', 'cool_42a', 'fake4b']

>>> [w for w in words if re.search()] ##### add your solution here
['hi42bye', 'nice1423', 'cool_42a']
```

d) For the given input list, filter all elements that start with `den` or end with `ly`.

```
>>> items = ['lovely', '1\ndentist', '2 lonely', 'eden', 'fly\n', 'dent']

>>> [e for e in items if ] ##### add your solution here
['lovely', '2 lonely', 'dent']
```

e) For the given input string, change whole word `mall` to `1234` only if it is at the start of a line.

```
>>> para = '''\
... ball fall wall tall
... mall call ball pall
... wall mall ball fall
...allet wallet malls'''

>>> print(re.sub()) ##### add your solution here
ball fall wall tall
1234 call ball pall
wall mall ball fall
allet wallet malls
```

f) For the given list, filter all elements having a line starting with `den` or ending with `ly`.

```
>>> items = ['lovely', '1\ndentist', '2 lonely', 'eden', 'fly\nfar', 'dent']

##### add your solution here
['lovely', '1\ndentist', '2 lonely', 'fly\nfar', 'dent']
```

g) For the given input list, filter all whole elements `12\nthree` irrespective of case.

```
>>> items = ['12\nthree\n', '12\nThree', '12\nthree\n4', '12\nthree']
##### add your solution here
['12\nThree', '12\nthree']
```

h) For the given input list, replace `hand` with `X` for all elements that start with `hand` followed by at least one word character.

```
>>> items = ['handed', 'hand', 'handy', 'unhanded', 'handle', 'hand-2']
```

```
##### add your solution here
```

```
['Xed', 'hand', 'Xy', 'unhanded', 'Xle', 'hand-2']
```

i) For the given input list, filter all elements starting with `h` . Additionally, replace `e` with `X` for these filtered elements.

```
>>> items = ['handed', 'hand', 'handy', 'unhanded', 'handle', 'hand-2']
```

```
##### add your solution here
```

```
['handXd', 'hand', 'handy', 'handlX', 'hand-2']
```

Alternation and Grouping

Similar to logical OR, alternation in regular expressions allows you to combine multiple patterns. These patterns can have some common elements between them, in which case grouping helps to form terser expressions. This chapter will also discuss the precedence rules used to determine which alternation wins.

Alternation

A conditional expression combined with logical OR evaluates to `True` if any of the condition is satisfied. Similarly, in regular expressions, you can use `|` metacharacter to combine multiple patterns to indicate logical OR. The matching will succeed if any of the alternate pattern is found in the input string. These alternatives have the full power of a regular expression, for example they can have their own independent anchors. Here's some examples.

```
# match either 'cat' or 'dog'
>>> bool(re.search(r'cat|dog', 'I like cats'))
True
>>> bool(re.search(r'cat|dog', 'I like dogs'))
True
>>> bool(re.search(r'cat|dog', 'I like parrots'))
False

# replace either 'cat' at start of string or 'cat' at end of word
>>> re.sub(r'\Acat|cat\b', 'X', 'catapults concatenate cat scat')
'Xapults concatenate X sX'

# replace either 'cat' or 'dog' or 'fox' with 'mammal'
>>> re.sub(r'cat|dog|fox', 'mammal', 'cat dog bee parrot fox')
'mammal mammal bee parrot mammal'
```

You might infer from above examples that there can be cases where many alternations are required. The `join` string method can be used to build the alternation list automatically from an iterable of strings.

```
>>> '|'.join(['car', 'jeep'])
'car|jeep'

>>> words = ['cat', 'dog', 'fox']
>>> '|'.join(words)
'cat|dog|fox'

>>> re.sub('|'.join(words), 'mammal', 'cat dog bee parrot fox')
'mammal mammal bee parrot mammal'
```



In the above examples with `join` method, the elements do not contain any special regular expression characters. Strings having metacharacters will be discussed in [re.escape](#) section.

Grouping

Often, there are some common things among the alternatives. It could be common characters or qualifiers like the anchors. In such cases, you can group them using a pair of parentheses metacharacters. Similar to $a(b+c)d = abd+acd$ in maths, you get $a(b|c)d = abd|acd$ in regular expressions.

```
# without grouping
>>> re.sub(r'reform|rest', 'X', 'red reform read arrest')
'red X read arX'
# with grouping
>>> re.sub(r're(form|st)', 'X', 'red reform read arrest')
'red X read arX'

# without grouping
>>> re.sub(r'\bpar\b|\bpart\b', 'X', 'par spare part party')
'X spare X party'
# taking out common anchors
>>> re.sub(r'\b(par|part)\b', 'X', 'par spare part party')
'X spare X party'
# taking out common characters as well
# you'll later learn a better technique instead of using empty alternate
>>> re.sub(r'\bpar(|t)\b', 'X', 'par spare part party')
'X spare X party'
```



There's plenty more features to grouping than just forming terser RE. It will be discussed as they become relevant in coming chapters.

For now, this is a good place to show how to incorporate normal strings (from a variable, expression result, etc) while building a regular expression. For example, adding anchors to alternation list created using the `join` method.

```
>>> words = ['cat', 'par']
>>> '|'.join(words)
'cat|par'
# without word boundaries, any matching portion will be replaced
>>> re.sub('|'.join(words), 'X', 'cater cat concatenate par spare')
'Xer X conXenate X sXe'

# note how raw string is used on either side of concatenation
# avoid f-strings unless you know how to compensate for RE
>>> alt = re.compile(r'\b(' + '|'.join(words) + r')\b')
# only whole words will be replaced now
>>> alt.sub('X', 'cater cat concatenate par spare')
'cater X concatenate X spare'

# this is how the above RE looks as a normal string
>>> alt.pattern
'\\b(cat|par)\\b'
```

```
>>> alt.pattern == r'\b(cat|par)\b'
True
```

In the above example, you had to concatenate strings to add word boundaries. If you needed to add string anchors so that the pattern only matches whole string, you can use `re.fullmatch` instead of manually adding the anchors.

```
>>> terms = ['no', 'ten', 'it']
>>> items = ['dip', 'nobody', 'it', 'oh', 'no', 'bitten']

>>> pat = re.compile('|'.join(terms))

# matching only whole elements
>>> [w for w in items if(pat.fullmatch(w))]
['it', 'no']
# matching anywhere
>>> [w for w in items if(pat.search(w))]
['nobody', 'it', 'no', 'bitten']
```

Precedence rules

There's some tricky situations when using alternation. If it is used for testing a match to get `True/False` against a string input, there is no ambiguity. However, for other things like string replacement, it depends on a few factors. Say, you want to replace either `are` or `spared` — which one should get precedence? The bigger word `spared` or the substring `are` inside it or based on something else?

In Python, the alternative which matches earliest in the input string gets precedence. `re.Match` output comes handy to illustrate this concept.

```
>>> words = 'lion elephant are rope not'

# span shows the start and end+1 index of matched portion
# match shows the text that satisfied the search criteria
>>> re.search(r'on', words)
<re.Match object; span=(2, 4), match='on'>
>>> re.search(r'ant', words)
<re.Match object; span=(10, 13), match='ant'>

# starting index of 'on' < index of 'ant' for given string input
# so 'on' will be replaced irrespective of order
# count optional argument here restricts no. of replacements to 1
>>> re.sub(r'on|ant', 'X', words, count=1)
'liX elephant are rope not'
>>> re.sub(r'ant|on', 'X', words, count=1)
'liX elephant are rope not'
```

What happens if alternatives match on same index? The precedence is then left to right in the order of declaration.

```
>>> mood = 'best years'
>>> re.search(r'year', mood)
<re.Match object; span=(5, 9), match='year'>
>>> re.search(r'years', mood)
<re.Match object; span=(5, 10), match='years'>

# starting index for 'year' and 'years' will always be same
# so, which one gets replaced depends on the order of alternation
>>> re.sub(r'year|years', 'X', mood, count=1)
'best Xs'
>>> re.sub(r'years|year', 'X', mood, count=1)
'best X'
```

Another example (without `count` restriction) to drive home the issue:

```
>>> words = 'ear xerox at mare part learn eye'

# this is going to be same as: r'ar'
>>> re.sub(r'ar|are|art', 'X', words)
'eX xerox at mXe pXt leXn eye'

# this is going to be same as: r'are|ar'
>>> re.sub(r'are|ar|art', 'X', words)
'eX xerox at mX pXt leXn eye'

# phew, finally this one works as needed
>>> re.sub(r'are|art|ar', 'X', words)
'eX xerox at mX pX leXn eye'
```

If you do not want substrings to sabotage your replacements, a robust workaround is to sort the alternations based on length, longest first.

```
>>> words = ['hand', 'handy', 'handful']

>>> alt = re.compile('|'.join(sorted(words, key=len, reverse=True)))
>>> alt.pattern
'handful|handy|hand'

>>> alt.sub('X', 'hands handful handed handy')
'Xs X Xed X'

# without sorting, alternation order will come into play
>>> re.sub('|'.join(words), 'X', 'hands handful handed handy')
'Xs Xful Xed Xy'
```



See also [regular-expressions: alternation](#) for more information regarding alternation and precedence rules in various regular expression implementations.

Cheatsheet and Summary

Note	Description
	multiple RE combined as conditional OR each alternative can have independent anchors
' '.join(iterable)	programmatically combine multiple RE
()	group pattern(s)
a(b c)d	same as abd acd
Alternation precedence	pattern which matches earliest in the input gets precedence tie-breaker is left to right if patterns have same starting location robust solution: sort the alternations based on length, longest first ' '.join(sorted(iterable, key=len, reverse=True))

So, this chapter was about specifying one or more alternate matches within the same RE using | metacharacter. Which can further be simplified using () grouping if the alternations have common aspects. Among the alternations, earliest matching pattern gets precedence. Left to right ordering is used as a tie-breaker if multiple alternations match starting from the same location. You also learnt ways to programmatically construct a RE.

Exercises

a) For the given input list, filter all elements that start with den or end with ly

```
>>> items = ['lovely', '1\ndentist', '2 lonely', 'eden', 'fly\n', 'dent']

##### add your solution here
['lovely', '2 lonely', 'dent']
```

b) For the given list, filter all elements having a line starting with den or ending with ly .

```
>>> items = ['lovely', '1\ndentist', '2 lonely', 'eden', 'fly\nfar', 'dent']

##### add your solution here
['lovely', '1\ndentist', '2 lonely', 'fly\nfar', 'dent']
```

c) For the given input strings, replace all occurrences of removed or reed or received or refused with X .

```
>>> s1 = 'creed refuse removed read'
>>> s2 = 'refused reed redo received'

>>> pat = re.compile() ##### add your solution here

>>> pat.sub('X', s1)
'cX refuse X read'
>>> pat.sub('X', s2)
'X X redo X'
```

d) For the given input strings, replace all matches from the list words with A .

```

>>> s1 = 'plate full of slate'
>>> s2 = "slated for later, don't be late"
>>> words = ['late', 'later', 'slated']

>>> pat = re.compile()          ##### add your solution here

>>> pat.sub('A', s1)
'pA full of sA'
>>> pat.sub('A', s2)
"A for A, don't be A"

```

e) Filter all whole elements from the input list `items` based on elements listed in `words` .

```

>>> items = ['slate', 'later', 'plate', 'late', 'slates', 'slated ']
>>> words = ['late', 'later', 'slated']

>>> pat = re.compile()          ##### add your solution here

##### add your solution here
['later', 'late']

```

Escaping metacharacters

This chapter will show how to match metacharacters literally, for manually as well as programmatically constructed patterns. You'll also learn about escape sequences supported by the `re` module.

Escaping with `\`

You have seen a few metacharacters and escape sequences that help to compose a RE. To match the metacharacters literally, i.e. to remove their special meaning, prefix those characters with a `\` (backslash) character. To indicate a literal `\` character, use `\\`. Assuming these are all part of raw string, not normal strings.

```
# even though ^ is not being used as anchor, it won't be matched literally
>>> bool(re.search(r'b^2', 'a^2 + b^2 - C*3'))
False
# escaping will work
>>> bool(re.search(r'b\^2', 'a^2 + b^2 - C*3'))
True

# match ( or ) literally
>>> re.sub(r'\(|\|)', '', '(a*b) + c')
'a*b + c'

# note that here input string is also a raw string
>>> re.sub(r'\\', '/', r'\learn\by\example')
'/learn/by/example'
```

As emphasized earlier, regular expressions is just another tool to process text. Some examples and exercises presented in this book can be solved using normal string methods as well. For real world use cases, ask yourself if regular expressions is needed at all?

```
>>> eqn = 'f*(a^b) - 3*(a^b)'

# straightforward search and replace, no need RE shenanigans
>>> eqn.replace('(a^b)', 'c')
'f*c - 3*c'
```

`re.escape`

Okay, what if you have a string variable that must be used to construct a RE — how to escape all the metacharacters? Relax, `re.escape` function has got you covered. No need to manually take care of all the metacharacters or worry about changes in future versions.

```
>>> expr = '(a^b)'
# print used here to show results similar to raw string
>>> print(re.escape(expr))
\\(a\\^b\\)
```

```
# replace only at end of string
>>> eqn = 'f*(a^b) - 3*(a^b)'
>>> re.sub(re.escape(expr) + r'\Z', 'c', eqn)
'f*(a^b) - 3*c'
```

Recall that in [Alternation](#) section, `join` was used to dynamically construct RE pattern from an iterable of strings. However, that didn't handle metacharacters. Here's how you can use `re.escape` so that the resulting pattern will match the strings from the input iterable literally.

```
# iterable of strings, assume alternation precedence sorting isn't needed
>>> terms = ['a_42', '(a^b)', '2|3']
# using 're.escape' and 'join' to construct the pattern
>>> pat1 = re.compile('|'.join(re.escape(s) for s in terms))
# using only 'join' to construct the pattern
>>> pat2 = re.compile('|'.join(terms))

>>> print(pat1.pattern)
a_42|\\(a\\^b\\)|2\\|3
>>> print(pat2.pattern)
a_42|(a^b)|2|3

>>> s = 'ba_423 (a^b)c 2|3 a^b'
>>> pat1.sub('X', s)
'bX3 Xc X a^b'
>>> pat2.sub('X', s)
'bXX (a^b)c X|X a^b'
```

Escape sequences

Certain characters like tab and newline can be expressed using escape sequences as `\t` and `\n` respectively. These are similar to how they are treated in normal string literals. However, `\b` is for word boundaries as seen earlier, whereas it stands for backspace character in normal string literals.

The full list is mentioned at the end of [docs.python: Regular Expression Syntax](#) section as `\a \b \f \n \N \r \t \u \U \v \x \\`. Do read the documentation for details as well as how it differs for byte data.

```
>>> re.sub(r'\t', ':', 'a\tb\tc')
'a:b:c'

>>> re.sub(r'\n', ' ', '1\n2\n3')
'1 2 3'
```



If an escape sequence is not defined, you'll get an error.

```
>>> re.search(r'\e', 'hello')
re.error: bad escape \e at position 0
```

You can also represent a character using hexadecimal escape of the format `\xNN` where `NN` are exactly two hexadecimal characters. If you represent a metacharacter using escapes, it will be treated literally instead of its metacharacter feature.

```
# \x20 is space character
>>> re.sub(r'\x20', '', 'h e l l o')
'hello'

# \x7c is '|' character
>>> re.sub(r'2\x7c3', '5', '12|30')
'150'
>>> re.sub(r'2|3', '5', '12|30')
'15|50'
```



See [ASCII code table](#) for a handy cheatsheet with all the ASCII characters and their hexadecimal representation.

Octal escapes will be discussed in [Backreference](#) section. [Codepoints and Unicode escapes](#) section will discuss escapes for unicode characters using `\u` and `\U`.

Cheatsheet and Summary

Note	Description
<code>\</code>	prefix metacharacters with <code>\</code> to match them literally
<code>\\</code>	to match <code>\</code> literally
<code>re.escape</code>	automatically escape all metacharacters ex: <code>' '.join(re.escape(s) for s in iterable)</code>
<code>\t</code>	escape sequences like those supported in string literals
<code>\b</code>	word boundary in RE but backspace in string literals
<code>\e</code>	undefined escapes will result in an error
<code>\xNN</code>	represent a character using hexadecimal value
<code>\x7c</code>	will match <code> </code> literally

This short chapter discussed how to match metacharacters literally. `re.escape` helps if you are using input strings sourced from elsewhere to build the final RE. You also saw how to use escape sequences to represent characters and how they differ from normal string literals.

Exercises

a) Transform the given input strings to the expected output using same logic on both strings.

```
>>> str1 = '(9-2)*5+qty/3'
>>> str2 = '(qty+4)/2-(9-2)*5+pq/4'

##### add your solution here for str1
'35+qty/3'
```

```
##### add your solution here for str2
'(qty+4)/2-35+pq/4'
```

b) Replace `(4)\|` with `2` only at the start or end of given input strings.

```
>>> s1 = r'2.3/(4)\|6 foo 5.3-(4)\|'
>>> s2 = r'(4)\|42 - (4)\|3'
>>> s3 = 'two - (4)\|\n'

>>> pat = re.compile() ##### add your solution here

>>> pat.sub('2', s1)
'2.3/(4)\|6 foo 5.3-2'
>>> pat.sub('2', s2)
'242 - (4)\|3'
>>> pat.sub('2', s3)
'two - (4)\|\n'
```

c) Replace any matching element from the list `items` with `X` for given the input strings. Match the elements from `items` literally. Assume no two elements of `items` will result in any matching conflict.

```
>>> items = ['a.b', '3+n', r'x\y\z', 'qty||price', '{n}']
>>> pat = re.compile() ##### add your solution here

>>> pat.sub('X', '0a.bcd')
'0Xcd'
>>> pat.sub('X', 'E{n}AMPLE')
'EXAMPLE'
>>> pat.sub('X', r'43+n2 ax\y\ze')
'4X2 aXe'
```

d) Replace backspace character `\b` with a single space character for the given input string.

```
>>> ip = '123\b456'
>>> ip
'123\x08456'
>>> print(ip)
12456

>>> re.sub() ##### add your solution here
'123 456'
```

e) Replace all occurrences of `\e` with `e`.

```
>>> ip = r'th\er\e ar\e common asp\ects among th\e alt\ernations'

>>> re.sub() ##### add your solution here
'there are common aspects among the alternations'
```

f) Replace any matching item from the list `eqns` with `X` for given the string `ip`. Match the items from `eqns` literally.

```
>>> ip = '3-(a^b)+2*(a^b)-(a/b)+3'
>>> eqns = ['(a^b)', '(a/b)', '(a^b)+2']

##### add your solution here

>>> pat.sub('X', ip)
'3-X*X-X+3'
```

Dot metacharacter and Quantifiers

This chapter introduces dot metacharacter and metacharacters related to quantifiers. As the name implies, quantifiers allows you to specify how many times a character or grouping should be matched. With `*` string operator, you can do something like `'no' * 5` to get `'nonononono'`. This saves you manual repetition as well as gives the ability to programmatically repeat a string object as many times as you need. Quantifiers support this simple repetition as well as ways to specify a range of repetition. This range has the flexibility of being bounded or unbounded with respect to start and end values. Combined with dot metacharacter (and alternation if needed), quantifiers allow you to construct conditional AND logic between patterns.

Dot metacharacter

The dot metacharacter serves as a placeholder to match any character except the newline character.

```
# matches character 'c', any character and then character 't'
>>> re.sub(r'c.t', 'X', 'tac tin cat abc;tuv acute')
'taXin X abXuv aXe'

# matches character 'r', any two characters and then character 'd'
>>> re.sub(r'r..d', 'X', 'breadth markedly reported overrides')
'bXth maXly repoX oveXes'

# matches character '2', any character and then character '3'
>>> re.sub(r'2.3', '8', '42\t35')
'485'

# by default, dot metacharacter doesn't match newline character
>>> bool(re.search(r'a.b', 'a\nb'))
False
```

See [re.DOTALL](#) section to know how `.` metacharacter can match newline as well. Chapter [Character class](#) will discuss how to define your own custom placeholder for limited set of characters.

re.split

This chapter will additionally use `re.split` function to illustrate examples. For normal strings, you'd use the `str.split` method. For regular expressions, use the `re.split` function, whose argument list is shown below.

```
re.split(pattern, string, maxsplit=0, flags=0)
```

The first argument is the RE pattern to be used for splitting the input string, which is the second argument. `maxsplit` and `flags` are optional arguments. Output is a list of strings.

```
# same as: 'apple-85-mango-70'.split('-')
>>> re.split(r'-', 'apple-85-mango-70')
['apple', '85', 'mango', '70']

# maxsplit determines the maximum number of times to split the input
>>> re.split(r'-', 'apple-85-mango-70', maxsplit=1)
['apple', '85-mango-70']

# example with dot metacharacter
>>> re.split(r'::', 'bus:3:car:5:van')
['bus', 'car', 'van']
```

See [re.split with capture groups](#) section for details of how capture groups affect the output of `re.split` .

Greedy quantifiers

Quantifiers have functionality like the string repetition operator and range function. They can be applied to characters and groupings (and more, as you'll see in later chapters). Apart from the ability to specify exact quantity and bounded range, these can also match unbounded varying quantities. If the input string can satisfy a pattern with varying quantities in multiple ways, you can choose among two types of quantifiers to narrow down possibilities. In this section, **greedy** type of quantifiers is covered.

First up, the `?` metacharacter which quantifies a character or group to match `0` or `1` times. In other words, you make that character or group as something to be optionally matched. This leads to a terser RE compared to alternation and grouping.

```
# same as: r'ear|ar'
>>> re.sub(r'e?ar', 'X', 'far feat flare fear')
'fX feat flXe fX'

# same as: r'\bpar(t|)\b'
>>> re.sub(r'\bpart?\b', 'X', 'par spare part party')
'X spare X party'

# same as: r'\b(re.d|red)\b'
>>> words = ['red', 'read', 'ready', 're;d', 'road', 'redo', 'reed', 'rod']
>>> [w for w in words if re.search(r'\bre.?d\b', w)]
['red', 'read', 're;d', 'reed']

# same as: r'part|parrot'
>>> re.sub(r'par(ro)?t', 'X', 'par part parrot parent')
'par X X parent'

# same as: r'part|parent|parrot'
>>> re.sub(r'par(en|ro)?t', 'X', 'par part parrot parent')
'par X X X'
```

The `*` metacharacter quantifies a character or group to match `0` or more times. There is no upper bound.

```
# match 't' followed by zero or more of 'a' followed by 'r'
>>> re.sub(r'ta*r', 'X', 'tr tear tare steer sitaara')
'X tear Xe steer siXa'
# match 't' followed by zero or more of 'e' or 'a' followed by 'r'
>>> re.sub(r't(e|a)*r', 'X', 'tr tear tare steer sitaara')
'X X Xe sX siXa'
# match zero or more of '1' followed by '2'
>>> re.sub(r'1*2', 'X', '311111111125111142')
'3X511114X'
```

Here's some more examples with `re.split` function.

```
# last element is empty because there is nothing after 2 at the end of string
>>> re.split(r'1*2', '311111111125111142')
['3', '511114', '']
# later, you'll see how maxsplit helps to get behavior like str.partition
>>> re.split(r'1*2', '311111111125111142', maxsplit=1)
['3', '5111142']

# empty string matches at start and end of string
# it matches between every character
# and, there is an empty match after the split at u
>>> re.split(r'u*', 'cloudy')
['', 'c', 'l', 'o', '', 'd', 'y', '']
```

The `+` metacharacter quantifies a character or group to match `1` or more times. Similar to `*` quantifier, there is no upper bound. More importantly, this doesn't have surprises like matching empty string in between patterns or at the start/end of string.

```
>>> re.sub(r'ta+r', 'X', 'tr tear tare steer sitaara')
'tr tear Xe steer siXa'
>>> re.sub(r't(e|a)+r', 'X', 'tr tear tare steer sitaara')
'tr X Xe sX siXa'

>>> re.sub(r'1+2', 'X', '311111111125111142')
'3X5111142'

>>> re.split(r'1+', '311111111125111142')
['3', '25', '42']
>>> re.split(r'u+', 'cloudy')
['clo', 'dy']
```

You can specify a range of integer numbers, both bounded and unbounded, using `{}` metacharacters. There are four ways to use this quantifier as shown below:

Pattern	Description
<code>{m,n}</code>	match <code>m</code> to <code>n</code> times
<code>{m,}</code>	match at least <code>m</code> times
<code>{,n}</code>	match up to <code>n</code> times (including <code>0</code> times)
<code>{n}</code>	match exactly <code>n</code> times

```
>>> demo = ['abc', 'ac', 'adc', 'abbc', 'xabbbcz', 'bbb', 'bc', 'abbbbbc']

>>> [w for w in demo if re.search(r'ab{1,4}c', w)]
['abc', 'abbc', 'xabbbcz']
>>> [w for w in demo if re.search(r'ab{3,}c', w)]
['xabbbcz', 'abbbbbc']
>>> [w for w in demo if re.search(r'ab{,2}c', w)]
['abc', 'ac', 'abbc']
>>> [w for w in demo if re.search(r'ab{3}c', w)]
['xabbbcz']
```



The `{}` metacharacters have to be escaped to match them literally. However, unlike `()` metacharacters, these have lot more leeway. For example, escaping `{` alone is enough, or if it doesn't conform strictly to any of the four forms listed above, escaping is not needed at all.

Conditional AND

Next up, how to construct conditional AND using dot metacharacter and quantifiers.

```
# match 'Error' followed by zero or more characters followed by 'valid'
>>> bool(re.search(r'Error.*valid', 'Error: not a valid input'))
True

>>> bool(re.search(r'Error.*valid', 'Error: key not found'))
False
```

To allow matching in any order, you'll have to bring in alternation as well. That is somewhat manageable for 2 or 3 patterns. See [Conditional AND with lookarounds](#) section for an easier approach.

```
>>> seq1 = 'cat and dog'
>>> seq2 = 'dog and cat'
>>> bool(re.search(r'cat.*dog|dog.*cat', seq1))
True
>>> bool(re.search(r'cat.*dog|dog.*cat', seq2))
True

# if you just need True/False result, this would be a scalable approach
>>> patterns = (r'cat', r'dog')
>>> all(re.search(p, seq1) for p in patterns)
True
>>> all(re.search(p, seq2) for p in patterns)
True
```

What does greedy mean?

When you are using the `?` quantifier, how does Python decide to match `0` or `1` times, if both quantities can satisfy the RE? For example, consider the expression `re.sub(r'f.?o', 'X', 'foot')` — should `foo` be replaced or `fo`? It will always replace `foo`, because these are **greedy** quantifiers, meaning they try to match as much as possible.

```
>>> re.sub(r'f.?o', 'X', 'foot')
'Xt'

# a more practical example
# prefix '<' with '\' if it is not already prefixed
# both '<' and '\<' will get replaced with '\<'
# note the use of raw string for all the three arguments
>>> print(re.sub(r'\\?<', r'\<', r'blah \< foo < bar \< blah < baz'))
blah \< foo \< bar \< blah \< baz

# say goodbye to r'handful|handy|hand' shenanigans
>>> re.sub(r'hand(y|ful)?', 'X', 'hand handy handful')
'X X X'
```

But wait, how did the `r'Error.*valid'` example work? Shouldn't `.*` consume all the characters after `Error`? Good question. The regular expression engine actually does consume all the characters. Then realizing that the match failed, it gives back one character from end of string and checks again if the overall RE is satisfied. This process is repeated until a match is found or failure is confirmed. In regular expression parlance, this is called **backtracking**.

```
>>> sentence = 'that is quite a fabricated tale'

# r't.*a' will always match from first 't' to last 'a'
# also, note that count argument is set to 1 for illustration purposes
>>> re.sub(r't.*a', 'X', sentence, count=1)
'Xle'
>>> re.sub(r't.*a', 'X', 'star', count=1)
'sXr'

# matching first 't' to last 'a' for t.*a won't work for these cases
# the engine backtracks until .*q matches and so on
>>> re.sub(r't.*a.*q.*f', 'X', sentence, count=1)
'Xabricated tale'
>>> re.sub(r't.*a.*u', 'X', sentence, count=1)
'Xite a fabricated tale'
```



Backtracking can be quite time consuming for certain corner cases. Or even catastrophic — see [cloudflare: Details of the Cloudflare outage on July 2, 2019](#).

Non-greedy quantifiers

As the name implies, these quantifiers will try to match as minimally as possible. Also known as **lazy** or **reluctant** quantifiers. Appending a `?` to greedy quantifiers makes them non-greedy.

```
>>> re.sub(r'f.??o', 'X', 'foot', count=1)
'Xot'
>>> re.sub(r'f.??o', 'X', 'frost', count=1)
'Xst'

>>> re.sub(r'.{2,5}?', 'X', '123456789', count=1)
'X3456789'

>>> re.split(r':.*?:', 'green:3.14:teal::brown:oh!:blue')
['green', 'teal', 'brown', 'blue']
```

Like greedy quantifiers, lazy quantifiers will try to satisfy the overall RE.

```
>>> sentence = 'that is quite a fabricated tale'

# r't.*?a' will always match from first 't' to first 'a'
>>> re.sub(r't.*?a', 'X', sentence, count=1)
'Xt is quite a fabricated tale'

# matching first 't' to first 'a' for t.*?a won't work for this case
# so, engine will move forward until .*?f matches and so on
>>> re.sub(r't.*?a.*?f', 'X', sentence, count=1)
'Xabricated tale'
```

Cheatsheet and Summary

Note	Description
<code>.</code>	match any character except the newline character
greedy	match as much as possible
<code>?</code>	greedy quantifier, match <code>0</code> or <code>1</code> times
<code>*</code>	greedy quantifier, match <code>0</code> or more times
<code>+</code>	greedy quantifier, match <code>1</code> or more times
<code>{m,n}</code>	greedy quantifier, match <code>m</code> to <code>n</code> times
<code>{m,}</code>	greedy quantifier, match at least <code>m</code> times
<code>{,n}</code>	greedy quantifier, match up to <code>n</code> times (including <code>0</code> times)
<code>{n}</code>	greedy quantifier, match exactly <code>n</code> times
<code>pat1.*pat2</code>	any number of characters between <code>pat1</code> and <code>pat2</code>
<code>pat1.*pat2 pat2.*pat1</code>	match both <code>pat1</code> and <code>pat2</code> in any order
non-greedy	append <code>?</code> to greedy quantifier
	match as minimally as possible
<code>re.split</code>	split a string using regular expressions
	<code>re.split(pattern, string, maxsplit=0, flags=0)</code>
	<code>maxsplit</code> and <code>flags</code> are optional arguments

This chapter introduced the concept of specifying a placeholder instead of fixed string. When combined with quantifiers, you've seen a glimpse of how a simple RE can match wide range of text. In coming chapters, you'll learn how to create your own restricted set of placeholder characters.

Exercises



Since `.` metacharacter doesn't match newline character by default, assume that the input strings in the following exercises will not contain newline characters.

a) Replace `42//5` or `42/5` with `8` for the given input.

```
>>> ip = 'a+42//5-c pressure*3+42/5-14256'

>>> re.sub()      ##### add your solution here
'a+8-c pressure*3+8-14256'
```

b) For the list `items`, filter all elements starting with `hand` and ending with at most one more character or `le`.

```
>>> items = ['handed', 'hand', 'handled', 'handy', 'unhand', 'hands', 'handle']

##### add your solution here
['hand', 'handy', 'hands', 'handle']
```

c) Use `re.split` to get the output as shown for the given input strings.

```
>>> eqn1 = 'a+42//5-c'
>>> eqn2 = 'pressure*3+42/5-14256'
>>> eqn3 = 'r*42-5/3+42///5-42/53+a'

##### add your solution here for eqn1
['a+', '-c']
##### add your solution here for eqn2
['pressure*3+', '-14256']
##### add your solution here for eqn3
['r*42-5/3+42///5-', '3+a']
```

d) For the given input strings, remove everything from the first occurrence of `i` till end of the string.

```
>>> s1 = 'remove the special meaning of such constructs'
>>> s2 = 'characters while constructing'

>>> pat = re.compile()      ##### add your solution here

>>> pat.sub('', s1)
'remove the spec'
>>> pat.sub('', s2)
'characters wh'
```

e) For the given strings, construct a RE to get output as shown.

```
>>> str1 = 'a+b(addition)'  
>>> str2 = 'a/b(division) + c%d(#modulo)'  
>>> str3 = 'Hi there(greeting). Nice day(a(b))'  
  
>>> remove_parentheses = re.compile() ##### add your solution here  
  
>>> remove_parentheses.sub('', str1)  
'a+b'  
>>> remove_parentheses.sub('', str2)  
'a/b + c%d'  
>>> remove_parentheses.sub('', str3)  
'Hi there. Nice day'
```

f) Correct the given RE to get the expected output.

```
>>> words = 'plink incoming tint winter in caution sentient'  
>>> change = re.compile(r'int|in|ion|ing|inco|inter|ink')  
  
# wrong output  
>>> change.sub('X', words)  
'plXk XcomXg tX wXer X cautX sentient'  
  
# expected output  
>>> change = re.compile() ##### add your solution here  
>>> change.sub('X', words)  
'plX XmX tX wX X cautX sentient'
```

g) For the given greedy quantifiers, what would be the equivalent form using `{m,n}` representation?

- `?` is same as
- `*` is same as
- `+` is same as

h) `(a*|b*)` is same as `(a|b)*` — True or False?

i) For the given input strings, remove everything from first the occurrence of `test` (irrespective of case) till end of the string, provided `test` isn't at the end of the string.

```
>>> s1 = 'this is a Test'  
>>> s2 = 'always test your RE for corner cases'  
>>> s3 = 'a TEST of skill tests?'  
  
>>> pat = re.compile() ##### add your solution here  
  
>>> pat.sub('', s1)  
'this is a Test'  
>>> pat.sub('', s2)  
'always '  
>>> pat.sub('', s3)  
'a '
```

j) For the input list `words` , filter all elements starting with `s` and containing `e` and `t` in any order.

```
>>> words = ['sequoia', 'subtle', 'exhibit', 'asset', 'sets', 'tests', 'site']

##### add your solution here
['subtle', 'sets', 'site']
```

k) For the input list `words` , remove all elements having less than `6` characters.

```
>>> words = ['sequoia', 'subtle', 'exhibit', 'asset', 'sets', 'tests', 'site']

##### add your solution here
['sequoia', 'subtle', 'exhibit']
```

l) For the input list `words` , filter all elements starting with `s` or `t` and having a maximum of `6` characters.

```
>>> words = ['sequoia', 'subtle', 'exhibit', 'asset', 'sets', 'tests', 'site']

##### add your solution here
['subtle', 'sets', 'tests', 'site']
```

m) Can you reason out why this code results in the output shown? The aim was to remove all `<characters>` patterns but not the `<>` ones. The expected result was `'a 1<> b 2<> c'` .

```
>>> ip = 'a<apple> 1<> b<bye> 2<> c<cat>'

>>> re.sub(r'<.+?>', '', ip)
'a 1 2'
```

n) Use `re.split` to get the output as shown below for given input strings.

```
>>> s1 = 'go there // "this // that"'
>>> s2 = 'a//b // c//d e//f // 4//5'
>>> s3 = '42// hi//bye//see // carefully'

>>> pat = re.compile() ##### add your solution here

>>> pat.split() ##### add your solution here for s1
['go there', '"this // that"']
>>> pat.split() ##### add your solution here for s2
['a//b', 'c//d e//f // 4//5']
>>> pat.split() ##### add your solution here for s3
['42// hi//bye//see', 'carefully']
```

Interlude: Tools for debugging and visualization

As your RE gets complicated, it can get difficult to debug if you run into issues. Building your RE step by step from scratch and testing against input strings will go a long way in correcting the problem. To aid in such a process, you could use [various online tools](#).

regex101

[regex101](#) is a popular site to test your RE. You'll have to first choose the flavor as Python. Then you can add your RE, input strings, choose flags and an optional replacement string. Matching portions will be highlighted and explanation is offered in separate panes.

The below image is a screenshot from this link — [regex101](#): `r'ab{0,2}c'`

The screenshot shows the regex101 interface. The 'REGULAR EXPRESSION' field contains `r'ab{0,2}c'`. The 'TEST STRING' field contains `abc ac xabbbcz bc abbc abbbbbc`. The 'EXPLANATION' pane shows the following details:

- `" ab{0,2}c "` matches the character `a` literally (case sensitive)
- `b{0,2}` matches the character `b` literally (case sensitive)
- `{0,2}` Quantifier — Matches between `0` and `2` times, as many times as possible, giving back as needed (*greedy*)
- `c` matches the character `c` literally (case sensitive)
- Global pattern flags**
 - `g` modifier: global. All matches (don't return after first match)

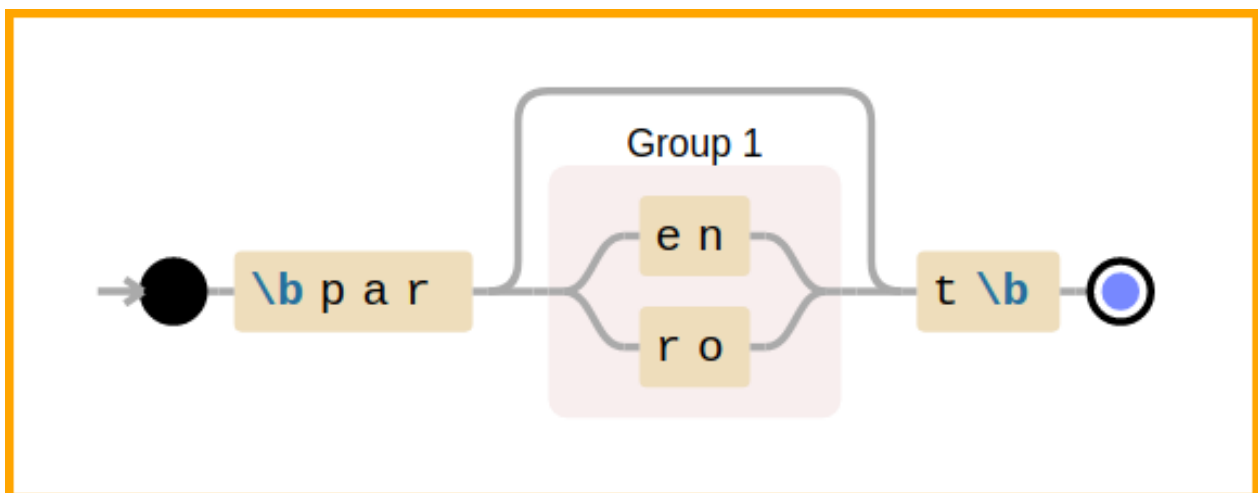


Do explore all the features provided by the site. There's a quick reference and other features like link sharing, code generator, quiz, cheatsheet, etc.

debuggex

Another useful tool is [debuggex](#) which converts your RE to a rail road diagram, thus providing a visual aid to understanding the pattern.

The below image is a screenshot from this link — [debuggex](#): `r'\bpar(en|ro)?t\b'`



regexcrossword

For practice, [regexcrossword](#) is often recommended. It only supports JavaScript, so some of the puzzles may not work the same with Python syntax. See [regexcrossword: howtoplay](#) for help.

The below image is a screenshot from this link — [regexcrossword: tutorial puzzle 5](#)

The screenshot displays the regexcrossword interface for a tutorial puzzle. At the top, a crossword grid is shown with two empty cells. The cell to the left of the top cell is labeled 'A|C', and the cell below it is labeled 'B'. The cell to the right of the top cell is labeled 'A?B?'. Below the grid, there are three buttons: a left arrow, a blue 'Validate' button, and a right arrow. At the bottom, there is a progress bar with numbers 1 through 9, where the number 5 is highlighted in blue. Below the progress bar is a 'Home' link.

Summary

This chapter briefly presented three online tools that can help you with understanding and interactively solving/debugging regular expressions. Syntax and features can vary, sometimes significantly, between various tools and programming languages. So, ensure that the program you are using supports the flavor of regular expressions you are using.

Working with matched portions

Having seen a few features that can match varying text, you'll learn how to extract and work with those matching portions in this chapter. First, you'll learn in detail about `re.Match` object. And then you'll learn about `re.findall` and `re.finditer` functions to get all the matches instead of just the first match. You'll also learn a few tricks like using functions in replacement section of `re.sub` and the use of `re.subn` function.

re.Match object

The `re.search` and `re.fullmatch` functions return a `re.Match` object from which various details can be extracted like the matched portion of string, location of matched portion, etc. Note that you get the details only for the first match, you'll see multiple matches later in this chapter. Here's some example with `re.Match` output.

```
>>> re.search(r'ab*c', 'abc ac adc abbbc')
<re.Match object; span=(0, 3), match='abc'>

>>> re.fullmatch(r'1(2|3)*4', '1233224')
<re.Match object; span=(0, 7), match='1233224'>
```

The details in the output above are for quick reference only. There are methods and attributes that you can apply on the `re.Match` object to get only the exact information you need. Use `span` method to get the **starting** and **ending + 1** indexes of the matching portion.

```
>>> sentence = 'that is quite a fabricated tale'
>>> m = re.search(r'q.*?t', sentence)
>>> m.span()
(8, 12)
>>> m.span()[0]
8

# you can also directly use the method without using intermediate variable
>>> re.search(r'q.*?t', sentence).span()
(8, 12)
```

The `()` grouping is also known as a **capture group**. It has multiple uses, one of which is the ability to work with matched portions of those groups. When capture groups are used with `re.search` or `re.fullmatch`, they can be retrieved using an index or `group` method on the `re.Match` object. The first element is always the entire matched portion and rest of the elements are for capture groups if they are present. The leftmost `()` in the pattern will get group number `1`, second leftmost `()` will get group number `2` and so on. Use `groups` method to get a tuple of only the capture group portions.

```
>>> re.search(r'b.*d', 'abc ac adc abbbc')
<re.Match object; span=(1, 9), match='bc ac ad'>
# retrieving entire matched portion using index
>>> re.search(r'b.*d', 'abc ac adc abbbc')[0]
'bc ac ad'
# retrieving entire matched portion using 'group' method
```

```
# you can also skip passing '0' as that is the default value
>>> re.search(r'b.*d', 'abc ac adc abbbc').group(0)
'bc ac ad'

# capture group example
>>> m = re.fullmatch(r'a(?:) (.)d(.)c', 'abc ac adc abbbc')
# to get matched portion of second capture group, can also use m.group(2)
>>> m[2]
'ac a'
# to get matched portion of third and first capture groups
>>> m.group(3, 1)
('c abbb', 'bc')
# to get a tuple of all the capture groups
# note that this will not have the entire matched portion
>>> m.groups()
('bc', 'ac a', 'c abbb')
```

To get the matching locations for the capture groups, pass the group number to `span` method. You can also use `start` and `end` methods to get either of those locations. Passing `0` is optional when you need the information for the entire matched portion.

```
>>> m = re.search(r'w(.*?)me', 'awesome')

>>> m.span()
(1, 7)
>>> m.span(1)
(2, 5)

>>> m.start()
1
>>> m.end(1)
5
```



There are many more methods and attributes available. See [docs.python: Match Objects](https://docs.python.org/3/library/re.html#match-objects) for details.

```
>>> pat = re.compile(r'hi.*bye')
>>> m = pat.search('This is goodbye then', 1, 15)
>>> m.pos
1
>>> m.endpos
15
>>> m.re
re.compile('hi.*bye')
>>> m.string
'This is goodbye then'
```

`groupdict` method will be covered in [Named capture groups](#) section and `expand` method will be covered in [Match.expand](#) section.

Assignment expressions

Since Python 3.8 introduced assignment expressions, it has become easier to work with matched portions in conditional statements.

```
# print capture group content only if the pattern matches
>>> if m := re.search(r'(.*)s', 'oh!'):
...     print(m[1])
...
>>> if m := re.search(r'(.*)s', 'awesome'):
...     print(m[1])
...
awe
```

This comes up often when you are processing a text file and the instructions depend on which pattern matches.

```
>>> text = ['type: fruit', 'date: 2020/04/28']
>>> for ip in text:
...     if m := re.search(r'type: (.*)', ip):
...         print(m[1])
...     elif m := re.search(r'date: (.*)/(.*)/', ip):
...         print(f'month: {m[2]}, year: {m[1]}')
...
fruit
month: 04, year: 2020
```



Did you know that [PEP 572](#) uses `re` module as one of the examples for assignment expressions?

Using functions in replacement section

Functions can be used in replacement section of `re.sub` instead of a string. A `re.Match` object will be passed to the function as argument. In [Backreference](#) section, you'll also learn how to directly reference the matching portions in replacement section.

```
# m[0] will contain entire matched portion
# a^2 and b^2 for the two matches in this example
>>> re.sub(r'(a|b)\^2', lambda m: m[0].upper(), 'a^2 + b^2 - C*3')
'A^2 + B^2 - C*3'

>>> re.sub(r'2|3', lambda m: str(int(m[0])**2), 'a^2 + b^2 - C*3')
'a^4 + b^4 - C*9'
```

Note that the output of the function has to be a string, otherwise you'll get an error. You'll see more examples with `lambda` and user defined functions in coming sections (for example, see [Numeric ranges](#) section).

Using dict in replacement section

Using a function in replacement section, you can specify a `dict` variable to determine the replacement string based on the matched text.

```
# one to one mappings
>>> d = { '1': 'one', '2': 'two', '4': 'four' }
>>> re.sub(r'1|2|4', lambda m: d[m[0]], '9234012')
'9two3four0onetwo'

# if the matched text doesn't exist as a key, default value will be used
# you'll later learn a much easier way to specify all digits
>>> re.sub(r'0|1|2|3|4|5|6|7|8|9', lambda m: d.get(m[0], 'X'), '9234012')
'XtwoXfourXonetwo'
```

For swapping two or more portions without using intermediate result, using a `dict` object is recommended.

```
>>> swap = { 'cat': 'tiger', 'tiger': 'cat' }
>>> words = 'cat tiger dog tiger cat'

>>> re.sub(r'cat|tiger', lambda m: swap[m[0]], words)
'tiger cat dog cat tiger'
```

For `dict` objects that have many entries and likely to undergo changes during development, building alternation list manually is not a good choice. Also, recall that as per precedence rules, longest length string should come first.

```
# note that numbers have been converted to strings here
# otherwise, you'd need to convert it in the lambda code
>>> d = { 'hand': '1', 'handy': '2', 'handful': '3', 'a^b': '4' }

# take care of metacharacter escaping first
>>> words = (re.escape(k) for k in d.keys())
# add anchors and flags if needed
>>> pat = re.compile('|'.join(sorted(words, key=len, reverse=True)))
>>> pat.pattern
'handful|handy|hand|a\\^b'
>>> pat.sub(lambda m: d[m[0]], 'handful hand pin handy (a^b)')
'3 1 pin 2 (4)'
```

re.findall

The `re.findall` function returns all the matched portions as a list of strings.

```
re.findall(pattern, string, flags=0)
```

The first argument is the RE pattern you want to test and extract against the input string, which is the second argument. `flags` is optional. Here's some examples.

```
>>> re.findall(r'ab*c', 'abc ac adc abbbc')
['abc', 'ac', 'abbbc']

>>> re.findall(r'ab+c', 'abc ac adc abbbc')
['abc', 'abbbc']

>>> s = 'PAR spar apparent SpArE part pare'
>>> re.findall(r'\bs?pare?\b', s, flags=re.I)
['PAR', 'spar', 'SpArE', 'pare']
```

It is useful for debugging purposes as well, for example to see the potential matches before applying substitution.

```
>>> re.findall(r't.*a', 'that is quite a fabricated tale')
['that is quite a fabricated ta']

>>> re.findall(r't.*?a', 'that is quite a fabricated tale')
['tha', 't is quite a', 'ted ta']
```

Presence of capture groups affects `re.findall` in different ways depending on number of groups used.

- If a single capture group is used, output will be a list of strings. Each element will have only the portion matched by the capture group
- If more than one capture group is used, output will be a list of tuples. Each element will be a tuple containing portions matched by all the capturing groups

For both cases, any pattern outside the capture groups will not be represented in the output. Also, you'll get an empty string if a particular capture group didn't match any character.

```
# without capture groups
>>> re.findall(r'ab*c', 'abc ac adc abbc xabbbcz bbb bc abbbbbc')
['abc', 'ac', 'abbc', 'abbbc', 'abbbbbc']

# with single capture group
>>> re.findall(r'a(b*)c', 'abc ac adc abbc xabbbcz bbb bc abbbbbc')
['b', '', 'bb', 'bbb', 'bbbbb']

# multiple capture groups
# note that last date didn't match because there's no comma at the end
# you'll later learn better ways to match such patterns
>>> re.findall(r'(.?)/(.?)/(.?)', '2020/04/25,1986/Mar/02,77/12/31')
[('2020', '04', '25'), ('1986', 'Mar', '02')]
```

See [Non-capturing groups](#) section if you need to use groupings without affecting `re.findall` output.

re.finditer

Use `re.finditer` to get an iterator object with each element as `re.Match` objects for each matched portion.

```
re.finditer(pattern, string, flags=0)
```

```
# output of finditer is an iterator object
>>> re.finditer(r'ab+c', 'abc ac adc abbbc')
<callable_iterator object at 0x7fb65e103438>

# each element is a re.Match object corresponding to the matched portion
>>> m_iter = re.finditer(r'ab+c', 'abc ac adc abbbc')
>>> for m in m_iter:
...     print(m)
...
<re.Match object; span=(0, 3), match='abc'>
<re.Match object; span=(11, 16), match='abbbc'>
```

Use the `re.Match` object's methods and attributes as needed. You can replicate `re.findall` functionality as well.

```
>>> m_iter = re.finditer(r'ab+c', 'abc ac adc abbbc')
>>> for m in m_iter:
...     print(m[0].upper(), m.span(), sep='\t')
...
ABC      (0, 3)
ABBBBC   (11, 16)

# same as: re.findall(r'(.*)/(.*)/(.*)', ' ', d)
>>> d = '2020/04/25,1986/Mar/02,77/12/31'
>>> m_iter = re.finditer(r'(.*)/(.*)/(.*)', ' ', d)
>>> [m.groups() for m in m_iter]
[('2020', '04', '25'), ('1986', 'Mar', '02')]
```



Since the output of `re.finditer` is an iterator object, you cannot iterate over it again without re-assigning. Not the case with `re.findall` which gives a list.

```
>>> d = '2020/04/25,1986/Mar/02,77/12/31'
>>> m_iter = re.finditer(r'(.*)', ' ', d)

>>> [m[1] for m in m_iter]
['2020/04/25', '1986/Mar/02']
>>> [m[1] for m in m_iter]
[]
```

re.split with capture groups

Capture groups affects `re.split` function as well. If the pattern used to split contains capture groups, the portions matched by those groups will also be a part of the output list.

```
# without capture group
>>> re.split(r'1*4?2', '3111111111251111426')
['3', '5', '6']

# to include the matching portions of the pattern as well in the output
>>> re.split(r'(1*4?2)', '3111111111251111426')
['3', '1111111112', '5', '111142', '6']
```

If part of the pattern is outside a capture group, the text thus matched won't be in the output. If a capture group didn't participate, it will be represented by `None` in the output list.

```
# here 4?2 is outside capture group, so that portion won't be in output
>>> re.split(r'(1*)4?2', '3111111111251111426')
['3', '111111111', '5', '1111', '6']

# multiple capture groups example
# note that the portion matched by b+ isn't present in the output
>>> re.split(r'(a+)b+(c+)', '3.14aabcccc42')
['3.14', 'aa', 'ccc', '42']

# here (4)? matches zero times on the first occasion
>>> re.split(r'(1*)(4)?2', '3111111111251111426')
['3', '111111111', None, '5', '1111', '4', '6']
```

Use of capture groups and `maxsplit=1` gives behavior similar to `str.partition` method.

```
# first element is portion before the first match
# second element is portion matched by the pattern itself
# third element is rest of the input string
>>> re.split(r'(a+b+c+)', '3.14aabcccc42abc88', maxsplit=1)
['3.14', 'aabccc', '42abc88']
```

re.subn

The `re.subn` has the same functionality as `re.sub` except that the output is a tuple. The first element of the tuple is the same output as `re.sub` function. The second element gives the number of substitutions made. In other words, you also get the number of matches.

```
re.subn(pattern, repl, string, count=0, flags=0)
```

```
>>> greeting = 'Have a nice weekend'

>>> re.sub(r'e', 'E', greeting)
'HavE a nicE wEEkEnd'

# with re.subn, you can infer that 5 substitutions were made
>>> re.subn(r'e', 'E', greeting)
('HavE a nicE wEEkEnd', 5)
```

Here's an example that performs conditional operation based on whether the substitution was successful.

```
>>> word = 'coffining'
# recursively delete 'fin'
>>> while True:
...     word, cnt = re.subn(r'fin', '', word)
...     if cnt == 0:
...         break
...
>>> word
'cog'
```

Cheatsheet and Summary

Note	Description
<code>re.Match</code> object	get details like matched portions, location, etc
<code>m[0]</code> or <code>m.group(0)</code>	entire matched portion of <code>re.Match</code> object <code>m</code>
<code>m[1]</code> or <code>m.group(1)</code>	matched portion of first capture group
<code>m[2]</code> or <code>m.group(2)</code>	matched portion of second capture group and so on
<code>m.groups()</code>	tuple of all the capture groups' matched portions
<code>m.span()</code>	start and end+1 index of entire matched portion
	pass a number to get span of that particular capture group
	can also use <code>m.start()</code> and <code>m.end()</code>
<code>re.sub(r'pat', f, s)</code>	function <code>f</code> will get <code>re.Match</code> object as argument
using <code>dict</code>	replacement string based on the matched text as dictionary key
	ex: <code>re.sub(r'pat', lambda m: d.get(m[0], default), s)</code>
<code>re.findall</code>	returns all the matches as a list of strings
	<code>re.findall(pattern, string, flags=0)</code>
	if 1 capture group is used, only its matches are returned
	1+, each element will be tuple of capture groups
	portion matched by pattern outside group won't be in output
	empty matches will be represented by empty string
<code>re.finditer</code>	iterator with <code>re.Match</code> object for each match
	<code>re.finditer(pattern, string, flags=0)</code>
<code>re.split</code>	capture groups affects <code>re.split</code> too
	text matched by the groups will be part of the output
	portion matched by pattern outside group won't be in output
	group that didn't match will be represented by <code>None</code>
<code>re.subn</code>	gives tuple of modified string and number of substitutions
	<code>re.subn(pattern, repl, string, count=0, flags=0)</code>

This chapter introduced different ways to work with various matching portions of input string. `re.Match` object helps you get the portion matched by the RE pattern and capture groups, location of the match, etc. Functions can be used in replacement section, which gets `re.Match` object as an argument. Using functions, you can do substitutions based on `dict` mappings. To get all the matches instead of just the first match, you can use `re.findall` (which gives

a list of strings as output) and `re.finditer` (which gives an iterator of `re.Match` objects). You also learnt how capture groups affect the output of `re.findall` and `re.split` functions. You'll see many more uses of groupings in coming chapters. The `re.subn` function is like `re.sub` but additionally gives number of matches as well.

Exercises

a) For the given strings, extract the matching portion from first `is` to last `t`.

```
>>> str1 = 'This the biggest fruit you have seen?'
>>> str2 = 'Your mission is to read and practice consistently'

>>> pat = re.compile() ##### add your solution here

##### add your solution here for str1
'is the biggest fruit'
##### add your solution here for str2
'ission is to read and practice consistent'
```

b) Find the starting index of first occurrence of `is` or `the` or `was` or `to` for the given input strings.

```
>>> s1 = 'match after the last newline character'
>>> s2 = 'and then you want to test'
>>> s3 = 'this is good bye then'
>>> s4 = 'who was there to see?'

>>> pat = re.compile() ##### add your solution here

##### add your solution here for s1
12
##### add your solution here for s2
4
##### add your solution here for s3
2
##### add your solution here for s4
4
```

c) Find the starting index of last occurrence of `is` or `the` or `was` or `to` for the given input strings.

```
>>> s1 = 'match after the last newline character'
>>> s2 = 'and then you want to test'
>>> s3 = 'this is good bye then'
>>> s4 = 'who was there to see?'

>>> pat = re.compile() ##### add your solution here

##### add your solution here for s1
12
##### add your solution here for s2
```

```

18
##### add your solution here for s3
17
##### add your solution here for s4
14

```

d) The given input string contains `:` exactly once. Extract all characters after the `:` as output.

```

>>> ip = 'fruits:apple, mango, guava, blueberry'

##### add your solution here
'apple, mango, guava, blueberry'

```

e) The given input strings contains some text followed by `-` followed by a number. Replace that number with its `log` value using `math.log()`.

```

>>> s1 = 'first-3.14'
>>> s2 = 'next-123'

>>> pat = re.compile() ##### add your solution here

>>> import math
>>> pat.sub() ##### add your solution here for s1
'first-1.144222799920162'
>>> pat.sub() ##### add your solution here for s2
'next-4.812184355372417'

```

f) Replace all occurrences of `par` with `spar`, `spare` with `extra` and `park` with `garden` for the given input strings.

```

>>> str1 = 'apartment has a park'
>>> str2 = 'do you have a spare cable'
>>> str3 = 'write a parser'

##### add your solution here

>>> pat.sub() ##### add your solution here for str1
'apartment has a garden'
>>> pat.sub() ##### add your solution here for str2
'do you have a extra cable'
>>> pat.sub() ##### add your solution here for str3
'write a sparser'

```

g) Extract all words between `(` and `)` from the given input string as a list. Assume that the input will not contain any broken parentheses.

```

>>> ip = 'another (way) to reuse (portion) matched (by) capture groups'

>>> re.findall() ##### add your solution here
['way', 'portion', 'by']

```

h) Extract all occurrences of `<` up to next occurrence of `>`, provided there is at least one

character in between `<` and `>` .

```
>>> ip = 'a<apple> 1< b<bye> 2< c<cat>'

>>> re.findall()          ##### add your solution here
['<apple>', '< b<bye>', '< c<cat>']
```

i) Use `re.findall` to get the output as shown below for the given input strings. Note the characters used in the input strings carefully.

```
>>> row1 = '-2,5 4,+3 +42,-53 4356246,-357532354 '
>>> row2 = '1.32,-3.14 634,5.63 63.3e3,9907809345343.235 '

>>> pat = re.compile()      ##### add your solution here

>>> pat.findall(row1)
[('-2', '5'), ('4', '+3'), ('+42', '-53'), ('4356246', '-357532354')]
>>> pat.findall(row2)
[('1.32', '-3.14'), ('634', '5.63'), ('63.3e3', '9907809345343.235')]
```

j) This is an extension to previous question.

- For `row1` , find the sum of integers of each tuple element. For example, sum of `-2` and `5` is `3` .
- For `row2` , find the sum of floating-point numbers of each tuple element. For example, sum of `1.32` and `-3.14` is `-1.82` .

```
>>> row1 = '-2,5 4,+3 +42,-53 4356246,-357532354 '
>>> row2 = '1.32,-3.14 634,5.63 63.3e3,9907809345343.235 '

# should be same as previous question
>>> pat = re.compile()      ##### add your solution here

##### add your solution here for row1
[3, 7, -11, -353176108]

##### add your solution here for row2
[-1.82, 639.63, 9907809408643.234]
```

k) Use `re.split` to get the output as shown below.

```
>>> ip = '42:no-output;1000:car-truck;SQEX49801'

>>> re.split()          ##### add your solution here
['42', 'output', '1000', 'truck', 'SQEX49801']
```

l) For the given list of strings, change the elements into a tuple of original element and number of times `t` occurs in that element.

```
>>> words = ['sequoia', 'attest', 'tattletale', 'asset']

##### add your solution here
[('sequoia', 0), ('attest', 3), ('tattletale', 4), ('asset', 1)]
```

m) The given input string has fields separated by `:`. Each field contains four uppercase alphabets followed optionally by two digits. Ignore the last field, which is empty. See [docs.python: Match.groups](#) and use `re.finditer` to get the output as shown below. If the optional digits aren't present, show `'NA'` instead of `None`.

```
>>> ip = 'TWXA42:JWPA:NTED01:'

##### add your solution here
[('TWXA', '42'), ('JWPA', 'NA'), ('NTED', '01')]
```



Note that this is different from `re.findall` which will just give empty string instead of `None` when a capture group doesn't participate.

n) Convert the comma separated strings to corresponding `dict` objects as shown below.

```
>>> row1 = 'name:rohan,maths:75,phy:89'
>>> row2 = 'name:rose,maths:88,phy:92'

>>> pat = re.compile() ##### add your solution here

##### add your solution here for row1
{'name': 'rohan', 'maths': '75', 'phy': '89'}
##### add your solution here for row2
{'name': 'rose', 'maths': '88', 'phy': '92'}
```

Character class

This chapter will discuss how to create your own custom placeholders to match limited set of characters and various metacharacters applicable inside character classes. You'll also learn about escape sequences for predefined character sets.

Custom character sets

Characters enclosed inside `[]` metacharacters is a character class (or set). It will result in matching any one of those characters once. It is similar to using single character alternations inside a grouping, but without the drawbacks of a capture group. In addition, character classes have their own versions of metacharacters and provide special predefined sets for common use cases. Quantifiers are applicable to character classes as well.

```
>>> words = ['cute', 'cat', 'cot', 'coat', 'cost', 'scuttle']

# same as: r'cot|cut' or r'c(o|u)t'
>>> [w for w in words if re.search(r'c[ou]t', w)]
['cute', 'cot', 'scuttle']

# same as: r'(a|e|o)+t'
>>> re.sub(r'[aeo]+t', 'X', 'meeting cute boat site foot')
'mXing cute bX site fX'
```

Range of characters

Character classes have their own metacharacters to help define the sets succinctly. Metacharacters outside of character classes like `^`, `$`, `()` etc either don't have special meaning or have completely different one inside the character classes. First up, the `-` metacharacter that helps to define a range of characters instead of having to specify them all individually.

```
# all digits
>>> re.findall(r'[0-9]+', 'Sample123string42with777numbers')
['123', '42', '777']

# whole words made up of lowercase alphabets and digits only
>>> re.findall(r'\b[a-z0-9]+\b', 'coat Bin food tar12 best')
['coat', 'food', 'tar12', 'best']

# whole words made up of lowercase alphabets, but starting with 'p' to 'z'
>>> re.findall(r'\b[p-z][a-z]*\b', 'coat tin food put stoop best')
['tin', 'put', 'stoop']

# whole words made up of only 'a' to 'f' and 'p' to 't' lowercase alphabets
>>> re.findall(r'\b[a-fp-t]+\b', 'coat tin food put stoop best')
['best']
```

Negating character sets

Next metacharacter is `^` which has to be specified as the first character of the character class. It negates the set of characters, so all characters other than those specified will be matched. As highlighted earlier, handle negative logic with care, you might end up matching more than you wanted.

```
# all non-digits
>>> re.findall(r'[^0-9]+', 'Sample123string42with777numbers')
['Sample', 'string', 'with', 'numbers']

# remove first two columns where : is delimiter
>>> re.sub(r'\A(?:[:]+){2}', '', 'foo:123:bar:baz', count=1)
'bar:baz'

# deleting characters at end of string based on a delimiter
>>> re.sub(r'=[^=]+\Z', '', 'foo=42; baz=123', count=1)
'foo=42; baz'

>>> dates = '2020/04/25,1986/Mar/02,77/12/31'
# Note that the third character set negates comma as well
# and comma is matched optionally outside the capture groups
>>> re.findall(r'([^\s]+)/([^\s]+)/([^\s,]+)?', dates)
[('2020', '04', '25'), ('1986', 'Mar', '02'), ('77', '12', '31')]
```

Sometimes, it is easier to use positive character class and negate the `re.search` result instead of using negated character class.

```
>>> words = ['tryst', 'fun', 'glyph', 'pity', 'why']

# words not containing vowel characters
>>> [w for w in words if re.search(r'\A[^aeiou]+\Z', w)]
['tryst', 'glyph', 'why']

# easier to write and maintain, note the 'not' logical operator
# but this'll match empty strings too unlike the previous solution
>>> [w for w in words if not re.search(r'[aeiou]', w)]
['tryst', 'glyph', 'why']
```

Matching metacharacters literally

Similar to other metacharacters, prefix `\` to character class metacharacters to match them literally. Some of them can be achieved by different placement as well.

- should be first or last character or escaped using `\`.

```
>>> re.findall(r'\b[a-z-]{2,}\b', 'ab-cd gh-c 12-423')
['ab-cd', 'gh-c']

>>> re.findall(r'\b[a-z\-\0-9]{2,}\b', 'ab-cd gh-c 12-423')
['ab-cd', 'gh-c', '12-423']
```

`^` should be other than first character or escaped using `\`.

```
>>> re.findall(r'a[+^]b', 'f*(a^b) - 3*(a+b)')
['a^b', 'a+b']

>>> re.findall(r'a[\\+^]b', 'f*(a^b) - 3*(a+b)')
['a^b', 'a+b']
```

`[]` can be escaped with `\` or placed as last character. `]` can be escaped with `\` or placed as first character.

```
>>> re.search(r'[]a-z0-9[]+', 'words[5] = tea')[0]
'words[5]'

>>> re.search(r'[a-z\\[]0-9[]+', 'words[5] = tea')[0]
'words[5]'
```

`\` should be escaped using `\`.

```
# note that the input string is using raw-string format
>>> print(re.search(r'[a\\b]+', r'5ba\babc2')[0])
ba\bab
```

Escape sequence sets

Commonly used character sets have predefined escape sequences:

- `\w` is similar to `[a-zA-Z0-9_]` for matching word characters (recall the definition for word boundaries)
- `\d` is similar to `[0-9]` for matching digit characters
- `\s` is similar to `[\t\n\r\f\v]` for matching whitespace characters

These escape sequences can be used as a standalone sequence or inside a character class.

```
>>> re.split(r'\d+', 'Sample123string42with777numbers')
['Sample', 'string', 'with', 'numbers']

>>> re.findall(r'\d+', 'foo=5, bar=3; x=83, y=120')
['5', '3', '83', '120']

>>> ''.join(re.findall(r'\b\w', 'sea eat car rat eel tea'))
'secret'

>>> re.findall(r'[\w\s]+', 'tea sea-pit sit-lean\tbean')
['tea sea', 'pit sit', 'lean\tbean']
```

And negative logic strikes again. Use `\W`, `\D`, and `\S` respectively for their negated sets.

```
>>> re.sub(r'\D+', '-', 'Sample123string42with777numbers')
'-123-42-777-'

>>> re.sub(r'\W+', '', 'foo=5, bar=3; x=83, y=120')
'foo5bar3x83y120'
```

```
# this output can be achieved with normal string method too, guess which one!
>>> re.findall(r'\S+', ' 1.3 \v\f foo_baz 42\tzzz \r\n1-2-3 ')
['1.3', 'foo_baz', '42', 'zzz', '1-2-3']
```



As mentioned before, the examples and description assume input made up of ASCII characters only, unless otherwise specified. These sets would behave differently depending on flags used. See [Unicode](#) chapter for more details.

Numeric ranges

Character classes can also be used to construct numeric ranges. However, it is easy to miss corner cases and some ranges are complicated to design.

```
# numbers between 10 to 29
>>> re.findall(r'\b[12]\d\b', '23 154 12 26 98234')
['23', '12', '26']

# numbers >= 100
>>> re.findall(r'\b\d{3,}\b', '23 154 12 26 98234')
['154', '98234']

# numbers >= 100 if there are leading zeros
>>> re.findall(r'\b0*[1-9]\d{2,}\b', '0501 035 154 12 26 98234')
['0501', '154', '98234']
```

If numeric range is difficult to construct, better to convert the matched portion to appropriate numeric format first.

```
# numbers < 350
>>> m_iter = re.finditer(r'\d+', '45 349 651 593 4 204')
>>> [m[0] for m in m_iter if int(m[0]) < 350]
['45', '349', '4', '204']

# note that return value is string and s[0] is used to get matched portion
>>> def num_range(s):
...     return '1' if 200 <= int(s[0]) <= 650 else '0'
...

# numbers between 200 and 650
# note that only function name is supplied, () is not used
# re.Match object is automatically passed as argument
>>> re.sub(r'\d+', num_range, '45 349 651 593 4 204')
'0 1 0 1 0 1'
```



See [regular-expressions: matching numeric ranges](#) for more examples.

Cheatsheet and Summary

Note	Description
<code>[ae;o]</code>	match any of these characters once quantifiers are applicable to character classes too
<code>[3-7]</code>	range of characters from <code>3</code> to <code>7</code>
<code>[^=b2]</code>	match other than <code>=</code> or <code>b</code> or <code>2</code>
<code>[a-z-]</code>	<code>-</code> should be first/last or escaped using <code>\</code> to match literally
<code>[+^]</code>	<code>^</code> shouldn't be first character or escaped using <code>\</code>
<code>[a\[\\]</code>	<code>[</code> can be escaped with <code>\</code> or placed as last character
<code>[a\[\\]</code>	<code>]</code> can be escaped with <code>\</code> or placed as first character
<code>[a\\b]</code>	<code>\</code> should be escaped using <code>\</code>
<code>\w</code>	similar to <code>[a-zA-Z0-9_]</code> for matching word characters
<code>\d</code>	similar to <code>[0-9]</code> for matching digit characters
<code>\s</code>	similar to <code>[\t\n\r\f\v]</code> for matching whitespace characters <code>\W</code> , <code>\D</code> , and <code>\S</code> for their opposites respectively

This chapter focused on how to create custom placeholders for limited set of characters. Grouping and character classes can be considered as two levels of abstractions. On the one hand, you can have character sets inside `[]` and on the other, you can have multiple alternations grouped inside `()` including character classes. As anchoring and quantifiers can be applied to both these abstractions, you can begin to see how regular expressions is considered a mini-programming language. In coming chapters, you'll even see how to negate groupings similar to negated character class in certain scenarios.

Exercises

a) For the list `items` , filter all elements starting with `hand` and ending with `s` or `y` or `le` .

```
>>> items = ['-handy', 'hand', 'handy', 'unhand', 'hands', 'handle']

##### add your solution here
['handy', 'hands', 'handle']
```

b) Replace all whole words `reed` or `read` or `red` with `X` .

```
>>> ip = 'redo red credible :read: rod reed'

##### add your solution here
'redo X credible :X: rod X'
```

c) For the list `words` , filter all elements containing `e` or `i` followed by `l` or `n` . Note that the order mentioned should be followed.

```
>>> words = ['surrender', 'unicorn', 'newer', 'door', 'empty', 'eel', 'pest']

##### add your solution here
['surrender', 'unicorn', 'eel']
```

d) For the list `words`, filter all elements containing `e` or `i` and `l` or `n` in any order.

```
>>> words = ['surrender', 'unicorn', 'newer', 'door', 'empty', 'eel', 'pest']

##### add your solution here
['surrender', 'unicorn', 'newer', 'eel']
```

e) Extract all hex character sequences, with `0x` optional prefix. Match the characters case insensitively, and the sequences shouldn't be surrounded by other word characters.

```
>>> str1 = '128A foo 0xfe32 34 0xbar'
>>> str2 = '0XDEADBEEF place 0x0fflce bad'

>>> hex_seq = re.compile() ##### add your solution here

##### add your solution here for str1
['128A', '0xfe32', '34']

##### add your solution here for str2
['0XDEADBEEF', '0x0fflce', 'bad']
```

f) Delete from `(` to the next occurrence of `)` unless they contain parentheses characters in between.

```
>>> str1 = 'def factorial()'
>>> str2 = 'a/b(division) + c%d(#modulo) - (e+(j/k-3)*4)'
>>> str3 = 'Hi there(greeting). Nice day(a(b))'

>>> remove_parentheses = re.compile() ##### add your solution here

>>> remove_parentheses.sub('', str1)
'def factorial'
>>> remove_parentheses.sub('', str2)
'a/b + c%d - (e*4)'
>>> remove_parentheses.sub('', str3)
'Hi there. Nice day(a'
```

g) For the list `words`, filter all elements not starting with `e` or `p` or `u`.

```
>>> words = ['surrender', 'unicorn', 'newer', 'door', 'empty', 'eel', 'pest']

##### add your solution here
['surrender', 'newer', 'door']
```

h) For the list `words`, filter all elements not containing with `u` or `w` or `ee` or `-`.

```
>>> words = ['p-t', 'you', 'tea', 'heel', 'owe', 'new', 'reed', 'ear']

##### add your solution here
['tea', 'ear']
```

i) The given input strings contain fields separated by `,` and fields can be empty too. Replace last three fields with `WHTSZ323`.

```
>>> row1 = '(2),kite,12,,D,C,, '
>>> row2 = 'hi,bye,sun,moon'

>>> pat = re.compile()          ##### add your solution here

>>> pat.sub()                    ##### add your solution here for row1
'(2),kite,12,,D,WHTSZ323'
>>> pat.sub()                    ##### add your solution here for row2
'hi,WHTSZ323'
```

j) Split the given strings based on consecutive sequence of digit or whitespace characters.

```
>>> str1 = 'lion \t Ink32onion Nice'
>>> str2 = '**1\f2\n3star\t7 77\r**'

>>> pat = re.compile()          ##### add your solution here

>>> pat.split(str1)
['lion', 'Ink', 'onion', 'Nice']
>>> pat.split(str2)
['**', 'star', '**']
```

k) Delete all occurrences of the sequence `<characters>` where `characters` is one or more non `>` characters and cannot be empty.

```
>>> ip = 'a<apple> 1< b<bye> 2< c<cat>'

##### add your solution here
'a 1< b 2< c'
```

l) `\b[a-z](on|no)[a-z]\b` is same as `\b[a-z][on]{2}[a-z]\b` . True or False? Sample input lines shown below might help to understand the differences, if any.

```
>>> print('known\nmood\nknow\npony\ninns')
known
mood
know
pony
inns
```

m) For the given list, filter all elements containing any number sequence greater than 624 .

```
>>> items = ['hi0000432abcd', 'car00625', '42_624 0512', '3.14 96 2 foo1234baz']

##### add your solution here
['car00625', '3.14 96 2 foo1234baz']
```

n) Count the maximum depth of nested braces for the given strings. Unbalanced or wrongly ordered braces should return `-1` . Note that this will require a mix of regular expressions and Python code.

```
>>> def max_nested_braces(ip):
##### add your solution here
```

```
>>> max_nested_braces('a*b')
0
>>> max_nested_braces('}a+b{')
-1
>>> max_nested_braces('a*b+{')
1
>>> max_nested_braces('{a+2}*{b+c}+e')
2
>>> max_nested_braces('{a+2}*{b+{c*d}}+e')
3
>>> max_nested_braces('{a+2}*{\n{b+{c*d}}+e*d}}')
4
>>> max_nested_braces('a*{b+c*{e*3.14}}}')
-1
```

o) By default, `str.split` method will split on whitespace and remove empty strings from the result. Which `re` module function would you use to replicate this functionality?

```
>>> ip = ' \t\r so pole\t\t\t\n\nlit in to \r\n\v\f '

>>> ip.split()
['so', 'pole', 'lit', 'in', 'to']
##### add your solution here
['so', 'pole', 'lit', 'in', 'to']
```

p) Convert the given input string to two different lists as shown below.

```
>>> ip = 'price_42 roast^\t\n^-ice==cat\neast'

##### add your solution here
['price_42', 'roast', 'ice', 'cat', 'east']

##### add your solution here
['price_42', ' ', 'roast', '^\t\n^-', 'ice', '==', 'cat', '\n', 'east']
```

q) Filter all elements whose first non-whitespace character is not a `#` character. Any element made up of only whitespace characters should be ignored as well.

```
>>> items = [' #comment', '\t\napple #42', '#oops', 'sure', 'no#1', '\t\r\f']

##### add your solution here
['\t\napple #42', 'sure', 'no#1']
```

Groupings and backreferences

This chapter will show how to reuse portion matched by capture groups via backreferences within RE definition and replacement section. You'll also learn some of the special grouping syntax for cases where plain capture groups isn't enough.

Backreference

Backreferences are like variables in a programming language. You have already seen how to use `re.Match` object to refer to the text captured by groups. Backreferences provide the same functionality, with the advantage that these can be directly used in RE definition as well as replacement section without having to invoke `re.Match` objects. Another advantage is that you can apply quantifiers to backreferences.

The syntax is `\N` or `\g<N>` where `N` is the capture group you want. The below syntax variations is applicable for **replacement section**, assuming they are used within raw strings.

- `\1` , `\2` up to `\99` to refer to the corresponding capture group
 - provided there are no digit characters after
 - `\0` and `\NNN` will be interpreted as octal value
- `\g<1>` , `\g<2>` etc (not limited to 99) to refer to the corresponding capture group
 - this also helps to avoid ambiguity between backreference and digits that follow
- `\g<0>` to refer to entire matched portion, similar to index `0` of `re.Match` objects
 - `\0` cannot be used because numbers starting with `0` are treated as octal value

Here's some examples with `\N` syntax.

```
# remove square brackets that surround digit characters
# note that use of raw strings for replacement string
>>> re.sub(r'\[(\d+)\]', r'\1', '[52] apples and [31] mangoes')
'52 apples and 31 mangoes'

# replace __ with _ and delete _ if it is alone
>>> re.sub(r'(_)?_', r'\1', '_foo_ __123__ _baz_')
'foo _123_ baz'

# swap words that are separated by a comma
>>> re.sub(r'(\w+),(\w+)', r'\2,\1', 'good,bad 42,24')
'bad,good 24,42'
```

Here's some examples with `\g<N>` syntax.

```
# ambiguity between \N and digit characters part of replacement string
>>> re.sub(r'\[(\d+)\]', r'(\15)', '[52] apples and [31] mangoes')
re.error: invalid group reference 15 at position 2
# \g<N> solves this issue
>>> re.sub(r'\[(\d+)\]', r'(\g<1>5)', '[52] apples and [31] mangoes')
'(525) apples and (315) mangoes'
# you can also use octal escapes
>>> re.sub(r'\[(\d+)\]', r'(\1\065)', '[52] apples and [31] mangoes')
'(525) apples and (315) mangoes'
```

```
# add something around the matched strings using \g<0>
>>> re.sub(r'[a-z]+', r'\g<0>{1}', '[52] apples and [31] mangoes')
'[52] {apples} {and} [31] {mangoes}'

# note the use of '+' instead of '*' quantifier to avoid empty matching
>>> re.sub(r'.+', r'Hi. \g<0>. Have a nice day', 'Hello world')
'Hi. Hello world. Have a nice day'

# duplicate first field and add it as last field
>>> re.sub(r'\A([^\,]+),.+', r'\g<0>,\1', 'fork,42,nice,3.14')
'fork,42,nice,3.14,fork'
```

Here's some examples for using backreferences within **RE definition**. Only `\N` syntax is available for use.

```
# whole words that have at least one consecutive repeated character
>>> words = ['effort', 'flee', 'facade', 'oddball', 'rat', 'tool']
>>> [w for w in words if re.search(r'\b\w*(\w)\1\w*\b', w)]
['effort', 'flee', 'oddball', 'tool']

# remove any number of consecutive duplicate words separated by space
# note the use of quantifier on backreferences
# use \W+ instead of space to cover cases like 'a;a-;a'
>>> re.sub(r'\b(\w+)(\1)+\b', r'\1', 'aa a a a 42 f_1 f_1 f_13.14')
'aa a 42 f_1 f_13.14'
```



Since `\g<N>` syntax is not available in RE definition, use octal escapes to avoid ambiguity between normal digit characters and backreferences.

```
>>> s = 'abcdefghijklmnaId'

# even though there's only one capture group, \11 will give an error
>>> re.sub(r'(.)*\11', 'X', s)
re.error: invalid group reference 11 at position 6
# use escapes for the digit portion to distinguish from the backreference
>>> re.sub(r'(.)*\1\x31', 'X', s)
'Xd'

# there are 12 capture groups here, so no error
# but requirement is \1 as backreference and 1 as normal digit
>>> re.sub(r'(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)*\11', 'X', s)
'abcdefghijklmnaId'
# use escapes again
>>> re.sub(r'(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)(.)*\1\x31', 'X', s)
'Xd'
```



It may be obvious, but it should be noted that backreference will provide the string that was matched, not the RE that was inside the capture group. For example, if `([0-9][a-f])` matches `3b`, then backreferencing will give `3b` and not any other valid match of RE like `8f`, `0a` etc. This is akin to how variables behave in programming, only the result of expression stays after variable assignment, not the expression itself. `regex` module supports [Subexpression calls](#) to refer to the RE itself.

Non-capturing groups

Grouping has many uses like applying quantifier on a RE portion, creating terse RE by factoring common portions and so on. It also affects behavior of functions like `re.findall` and `re.split` as seen in [Working with matched portions](#) chapter.

When backreferencing is not required, you can use a non-capturing group to avoid undesired behavior. It also helps to avoid keeping a track of capture group numbers when that particular group is not needed for backreferencing. The syntax is `(?:pat)` to define a non-capturing group. You'll see many more of such special groups starting with `(?` syntax later on.

```
# normal capture group will hinder ability to get whole match
# non-capturing group to the rescue
>>> re.findall(r'\b\w*(?:st|in)\b', 'cost akin more east run against')
['cost', 'akin', 'east', 'against']

# capturing wasn't needed here, only common grouping and quantifier
>>> re.split(r'hand(?:y|ful)?', '123hand42handy777handful500')
['123', '42', '777', '500']

# with normal grouping, need to keep track of all the groups
>>> re.sub(r'\A(([^,]+){3})([^\,]+)', r'\1(\3)', '1,2,3,4,5,6,7')
'1,2,3,(4),5,6,7'

# using non-capturing groups, only relevant groups have to be tracked
>>> re.sub(r'\A(?:[^\,]+){3})([^\,]+)', r'\1(\2)', '1,2,3,4,5,6,7')
'1,2,3,(4),5,6,7'
```

Referring to text matched by a capture group with a quantifier will give only the last match, not entire match. Use a capture group around the grouping and quantifier together to get the entire matching portion. In such cases, the inner grouping is an ideal candidate to use non-capturing group.

```
>>> s = 'hi 123123123 bye 456123456'
>>> re.findall(r'(123)+', s)
['123', '123']
>>> re.findall(r'(?:123)+', s)
['123123123', '123']
# note that this issue doesn't affect substitutions
>>> re.sub(r'(123)+', 'X', s)
'hi X bye 456X456'
```

```
>>> row = 'one,2,3.14,42,five'
# surround only fourth column with double quotes
# note the loss of columns in the first case
>>> re.sub(r'\A(?:[^\,]+){3}([^\,]+)', r'\1"\2"', row)
'3.14,"42",five'
>>> re.sub(r'\A(?:[^\,]+){3})([^\,]+)', r'\1"\2"', row)
'one,2,3.14,"42",five'
```

However, there are situations where capture groups cannot be avoided. In such cases, you'd need to manually work with `re.Match` objects to get desired results.

```
>>> words = 'effort flee facade oddball rat tool'
# whole words containing at least one consecutive repeated character
>>> repeat_char = re.compile(r'\b\w*(\w)\1\w*\b')

# () in findall will only return text matched by capture groups
>>> repeat_char.findall(words)
['f', 'e', 'l', 'o']
# finditer to the rescue
>>> m_iter = repeat_char.finditer(words)
>>> [m[0] for m in m_iter]
['effort', 'flee', 'oddball', 'tool']
```

Named capture groups

RE can get cryptic and difficult to maintain, even for seasoned programmers. There are a few constructs to help add clarity. One such is naming the capture groups and using that name for backreferencing instead of plain numbers. The syntax is `(?P<name>pat)` for naming the capture groups. The name used should be a valid Python identifier. Use `'name'` for `re.Match` objects, `\g<name>` in replacement section and `(?P=name)` for backreferencing in RE definition. These will still behave as normal capture groups, so `\N` or `\g<N>` numbering can be used as well.

```
# giving names to first and second captured words
>>> re.sub(r'(?P<fw>\w+),(?P<sw>\w+)', r'\g<sw>,\g<fw>', 'good,bad 42,24')
'bad,good 24,42'

>>> s = 'aa a a a 42 f_1 f_1 f_13.14'
>>> re.sub(r'\b(?P<dup>\w+)(?P=dup)+\b', r'\g<dup>', s)
'aa a 42 f_1 f_13.14'

>>> sentence = 'I bought an apple'
>>> m = re.search(r'(?P<fruit>\w+)\Z', sentence)
>>> m[1]
'apple'
>>> m['fruit']
'apple'
>>> m.group('fruit')
'apple'
```

You can use `groupdict` method on the `re.Match` object to extract the portions matched by named capture groups as a `dict` object. The capture group name will be the **key** and the portion matched by the group will be the **value**.

```
# single match
>>> details = '2018-10-25,car,2346'
>>> re.search(r'(?P<date>[^,]+),(?P<product>[^,]+)', details).groupdict()
{'date': '2018-10-25', 'product': 'car'}

# normal groups won't be part of the output
>>> re.search(r'(?P<date>[^,]+),([^,]+)', details).groupdict()
{'date': '2018-10-25'}

# multiple matches
>>> s = 'good,bad 42,24'
>>> [m.groupdict() for m in re.finditer(r'(?P<fw>\w+),(?P<sw>\w+)', s)]
[{'fw': 'good', 'sw': 'bad'}, {'fw': '42', 'sw': '24'}]
```

Conditional groups

This special grouping allows you to add a condition that depends on whether a capture group succeeded in matching. You can also add an optional else condition. The syntax as per the docs is shown below.

```
(?(id/name)yes-pattern|no-pattern)
```

Here `id` means the `N` used to backreference a capture group and `name` refers to the identifier used for a named capture group. Here's an example with **yes-pattern** alone being used. The task is to match elements containing word characters only or if it additionally starts with a double quote, it must end with a double quote.

```
>>> words = ['"hi"', 'bye', 'bad"', '"good"', '42', '"3']
>>> pat = re.compile(r'(")?\w+(?(1)"')
>>> [w for w in words if pat.fullmatch(w)]
['"hi"', 'bye', '"good"', '42']

# for this simple case, you can also expand it manually
# but for complex patterns, it is better to use conditional groups
# as it will avoid repeating the complex pattern
>>> [w for w in words if re.fullmatch(r'"?\w+"|\w+', w)]
['"hi"', 'bye', '"good"', '42']

# cannot simply use ? quantifier as they are independent, not constrained
>>> [w for w in words if re.fullmatch(r'"?\w+?"', w)]
['"hi"', 'bye', 'bad"', '"good"', '42', '"3']
# also, fullmatch plays a big role in avoiding partial match
>>> [w for w in words if pat.search(w)]
['"hi"', 'bye', 'bad"', '"good"', '42', '"3']
```

Here's an example with **no-pattern** as well.

```
# filter elements containing word characters surrounded by ()
# or, containing word characters separated by a hyphen
>>> words = ['(hi)', 'good-bye', 'bad', '(42)', '-oh', 'i-j', '(-)']

# same as: r'\(\w+\)|\w+-\w+'
>>> pat = re.compile(r'\(\)?\w+(?(1)\)|-\w+)')
>>> [w for w in words if pat.fullmatch(w)]
['(hi)', 'good-bye', '(42)', 'i-j']
```

Conditional groups have a very specific use case, and it is generally helpful for those cases. The main advantage is that it prevents pattern duplication, although that can also be achieved using [Subexpression calls](#) with `regex` module. Another advantage is that if the common pattern uses capture groups, then the duplication alternation method will need different back-reference numbers.

Match.expand

The `expand` method on `re.Match` objects accepts syntax similar to the replacement section of `re.sub` function. The difference is that `expand` method returns only the string after backreference expansion, instead of entire input string with modified content.

```
# re.sub vs Match.expand
>>> re.sub(r'w(.*?)m', r'[\1]', 'awesome')
'a[eso]e'
>>> re.search(r'w(.*?)m', 'awesome').expand(r'[\1]')
'[eso]'

# example with re.finditer
>>> dates = '2020/04/25,1986/03/02,77/12/31'
>>> m_iter = re.finditer(r'([^\s]+)/([^\s]+)/[^\s,?]', dates)
# same as: [f'Month:{m[2]}, Year:{m[1]}' for m in m_iter]
>>> [m.expand(r'Month:\2, Year:\1') for m in m_iter]
['Month:04, Year:2020', 'Month:03, Year:1986', 'Month:12, Year:77']
```

Cheatsheet and Summary

Note	Description
<code>\N</code>	backreference, gives matched portion of Nth capture group applies to both RE definition and replacement section possible values: <code>\1</code> , <code>\2</code> up to <code>\99</code> provided no more digits <code>\0</code> and <code>\NNN</code> will be treated as octal escapes
<code>\g<N></code>	backreference, gives matched portion of Nth capture group applies only to replacement section use octal escapes to prevent ambiguity in RE definition possible values: <code>\g<0></code> , <code>\g<1></code> , etc (not limited to 99) <code>\g<0></code> refers to entire matched portion

Note	Description
<code>(?:pat)</code>	non-capturing group
<code>(?P<name>pat)</code>	useful wherever grouping is required, but not backreference named capture group refer as <code>'name'</code> in <code>re.Match</code> object refer as <code>(?P=name)</code> in RE definition refer as <code>\g<name></code> in replacement section can also use <code>\N</code> and <code>\g<N></code> format if needed
<code>groupdict</code>	method applied on a <code>re.Match</code> object gives named capture group portions as a <code>dict</code>
<code>(?(id/name)yes no)</code>	conditional group match <code>yes-pattern</code> if backreferenced group succeeded else, match <code>no-pattern</code> which is optional
<code>expand</code>	method applied on a <code>re.Match</code> object accepts syntax like replacement section of <code>re.sub</code> gives back only string after backreference expansion

This chapter showed how to use backreferences to refer to portion matched by capture groups in both RE definition and replacement section. When capture groups results in unwanted behavior change (ex: `re.findall` and `re.split`), you can use non-capturing groups instead. Named capture groups add clarity to patterns and you can use `groupdict` method on a `re.Match` object to get a `dict` of matched portions. Conditional groups allows you to take an action based on another capture group succeeding or failing to match. There are more special groups to be discussed in coming chapters.

Exercises

a) Replace the space character that occurs after a word ending with `a` or `r` with a newline character.

```
>>> ip = 'area not a _a2_ roar took 22'

>>> print(re.sub())      ##### add your solution here
area
not a
_a2_ roar
took 22
```

b) Add `[]` around words starting with `s` and containing `e` and `t` in any order.

```
>>> ip = 'sequoia subtle exhibit asset sets tests site'

##### add your solution here
'sequoia [subtle] exhibit asset [sets] tests [site]'
```

c) Replace all whole words with `X` that start and end with the same word character. Single character word should get replaced with `X` too, as it satisfies the stated condition.

```
>>> ip = 'oreo not a _a2_ roar took 22'
```

```
##### add your solution here
```

```
'X not X X X took X'
```

d) Convert the given **markdown** headers to corresponding **anchor** tag. Consider the input to start with one or more `#` characters followed by space and word characters. The `name` attribute is constructed by converting the header to lowercase and replacing spaces with hyphens. Can you do it without using a capture group?

```
>>> header1 = '# Regular Expressions'
```

```
>>> header2 = '## Compiling regular expressions'
```

```
##### add your solution here for header1
```

```
'# <a name="regular-expressions"></a>Regular Expressions'
```

```
##### add your solution here for header2
```

```
'## <a name="compiling-regular-expressions"></a>Compiling regular expressions'
```

e) Convert the given **markdown** anchors to corresponding **hyperlinks**.

```
>>> anchor1 = '# <a name="regular-expressions"></a>Regular Expressions'
```

```
>>> anchor2 = '## <a name="subexpression-calls"></a>Subexpression calls'
```

```
##### add your solution here for anchor1
```

```
'[Regular Expressions](#regular-expressions)'
```

```
##### add your solution here for anchor2
```

```
'[Subexpression calls](#subexpression-calls)'
```

f) Count the number of whole words that have at least two occurrences of consecutive repeated alphabets. For example, words like `stillness` and `Committee` should be counted but not words like `root` or `readable` or `rotational`.

```
>>> ip = '''oppressed abandon accommodation bloodless
... carelessness committed apparition innkeeper
... occasionally afforded embarrassment foolishness
... depended successfully succeeded
... possession cleanliness suppress'''
```

```
##### add your solution here
```

```
13
```

g) For the given input string, replace all occurrences of digit sequences with only the unique non-repeating sequence. For example, `232323` should be changed to `23` and `897897` should be changed to `897`. If there no repeats (for example `1234`) or if the repeats end prematurely (for example `12121`), it should not be changed.

```
>>> ip = '1234 2323 453545354535 9339 11 60260260'
```

```
##### add your solution here
```

```
'1234 23 4535 9339 1 60260260'
```

h) Replace sequences made up of words separated by `:` or `.` by the first word of the sequence. Such sequences will end when `:` or `.` is not followed by a word character.

```
>>> ip = 'wow:Good:2_two:five: hi bye kite.777.water.'
```

```
##### add your solution here
'wow hi bye kite'
```

i) Replace sequences made up of words separated by `:` or `.` by the last word of the sequence. Such sequences will end when `:` or `.` is not followed by a word character.

```
>>> ip = 'wow:Good:2_two:five: hi bye kite.777.water.'
```

```
##### add your solution here
'five hi bye water'
```

j) Split the given input string on one or more repeated sequence of `cat` .

```
>>> ip = 'firecatlioncatcatcatbearcatcatparrot'
```

```
##### add your solution here
['fire', 'lion', 'bear', 'parrot']
```

k) For the given input string, find all occurrences of digit sequences with at least one repeating sequence. For example, `232323` and `897897` . If the repeats end prematurely, for example `12121` , it should not be matched.

```
>>> ip = '1234 2323 453545354535 9339 11 60260260'
```

```
>>> pat = re.compile() ##### add your solution here
```

```
# entire sequences in the output
```

```
##### add your solution here
```

```
['2323', '453545354535', '11']
```

```
# only the unique sequence in the output
```

```
##### add your solution here
```

```
['23', '4535', '1']
```

l) Convert the comma separated strings to corresponding `dict` objects as shown below. The keys are `name` , `maths` and `phy` for the three fields in the input strings.

```
>>> row1 = 'rohan,75,89'
```

```
>>> row2 = 'rose,88,92'
```

```
>>> pat = re.compile() ##### add your solution here
```

```
##### add your solution here for row1
```

```
{'name': 'rohan', 'maths': '75', 'phy': '89'}
```

```
##### add your solution here for row2
```

```
{'name': 'rose', 'maths': '88', 'phy': '92'}
```

m) Surround all whole words with `()` . Additionally, if the whole word is `imp` or `ant` , delete them. Can you do it with single substitution?

```
>>> ip = 'tiger imp goat eagle ant important'
```

```
##### add your solution here
```

```
'(tiger) () (goat) (eagle) () (important)'
```

n) Filter all elements that contains a sequence of lowercase alphabets followed by `-` followed by digits. They can be optionally surrounded by `{{` and `}}` . Any partial match shouldn't be part of the output.

```
>>> ip = ['{{apple-150}}', '{{mango2-100}}', '{{cherry-200}', 'grape-87']
```

```
##### add your solution here
```

```
['{{apple-150}}', 'grape-87']
```

o) The given input string has sequences made up of words separated by `:` or `.` and such sequences will end when `:` or `.` is not followed by a word character. For all such sequences, display only the last word followed by `-` followed by first word.

```
>>> ip = 'wow:Good:2_two:five: hi bye kite.777.water.'
```

```
##### add your solution here
```

```
['five-wow', 'water-kite']
```

Interlude: Common tasks

Tasks like matching phone numbers, ip addresses, dates, etc are so common that you can often find them collected as a library. This chapter shows some examples for [CommonRegex](#). `re` module documentation also has a section on tasks like [docs.python: tokenizer](#). See also [Awesome Regex: Collections](#).

CommonRegex

You can either install `commonregex` as a module or go through [commonregex.py](#) and choose the regular expression you need. There are several ways to use the patterns, see [Common-Regex: Usage](#) for details.

```
>>> from commonregex import ip

>>> data = 'hello 255.21.255.22 okay'
>>> ip.findall(data)
['255.21.255.22']
```

If you check `ip.pattern`, you'll find `(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)` repeated without any other restriction. So, this will not prevent partial matches. You'll have to add those conditions yourself.

```
>>> data = '23.14.2.4.2 255.21.255.22 567.12.2.1'

# wrong matches
>>> ip.findall(data)
['23.14.2.4', '255.21.255.22', '67.12.2.1']

# corrected usage
>>> [e for e in data.split() if ip.fullmatch(e)]
['255.21.255.22']
```

Summary

Some patterns are quite complex and not easy to build and validate from scratch. Libraries like **CommonRegex** are helpful to reduce your time and effort needed for commonly known tasks. However, you do need to **test** the solution for *your* use case. See also [stackoverflow: validating email addresses](#).

Lookarounds

You've already seen how to create custom character classes and various avatars of special groupings. In this chapter you'll learn more groupings, known as lookarounds, that help to create custom anchors and add conditions within RE definition. These assertions are also known as **zero-width patterns** because they add restrictions similar to anchors and are not part of the matched portions. Also, you will learn how to negate a grouping similar to negated character sets.

Conditional expressions

Before you get used to lookarounds too much, it is good to remember that Python is a programming language. You have control structures and you can combine multiple conditions using logical operators, functions like `all/any`, etc. Also, do not forget that `re` is only one of the tools available for string processing.

```
>>> items = ['1,2,3,4', 'a,b,c,d', '#foo 123']

# filter elements containing digit and '#' characters
>>> [s for s in items if re.search(r'\d', s) and '#' in s]
['#foo 123']

# modify elements only if it doesn't start with '#'
>>> for s in items:
...     if s[0] != '#':
...         print(re.sub(r',.+,', ' ', s))
...
1 4
a d
```

Negative lookarounds

Lookaround assertions can be added in two ways — **lookbehind** and **lookahead**. Each of these can be a **positive** or a **negative** assertion. Syntax wise, lookbehind has an extra `<` compared to the lookahead version. Negative lookarounds can be identified by the use of `!` whereas `=` is used for positive lookarounds. This section is about negative lookarounds, whose complete syntax is shown below.

- `(?!pat)` for negative lookahead assertion
- `(?<!pat)` for negative lookbehind assertion

As mentioned earlier, lookarounds are not part of matched portions and do not affect capture group numbering.

```
# change 'foo' only if it is not followed by a digit character
# note that end of string satisfies the given assertion
# 'foofoo' has two matches as the assertion doesn't consume characters
>>> re.sub(r'foo(?!\d)', 'baz', 'hey food! foo42 foot5 foofoo')
'hey bazd! foo42 bazt5 bazbaz'
```

```
# change 'foo' only if it is not preceded by _
# note how 'foo' at start of string is matched as well
>>> re.sub(r'(?<_!)foo', 'baz', 'foo _foo 42foofoo')
'baz _foo 42bazbaz'

# overlap example
# the final _ was replaced as well as played a part in the assertion
>>> re.sub(r'(?<_!)foo.', 'baz', 'food _fool 42foo_foot')
'baz _fool 42bazfoot'
```

Lookarounds can be mixed with already existing anchors and other features to define truly powerful restrictions:

```
# change whole word only if it is not preceded by : or -
>>> re.sub(r'(?<[::-])\b\w+\b', 'X', ':cart <apple -rest ;tea')
':cart <X -rest ;X'

# add space to word boundaries, but not at start or end of string
# similar to: re.sub(r'\b', ' ', 'foo_baz=num1+35*42/num2').strip()
>>> re.sub(r'(?<!\A)\b(?:\Z)', ' ', 'foo_baz=num1+35*42/num2')
'foo_baz = num1 + 35 * 42 / num2'
```

In all the examples so far, lookahead grouping was placed as a suffix and lookbehind as a prefix. This is how they are used most of the time, but not the only way to use them. Lookarounds can be placed anywhere and multiple lookarounds can be combined in any order. They do not consume characters nor do they play a role in matched portions. They just let you know whether the condition you want to test is satisfied from the current location in the input string.

```
# these two are equivalent
# replace a character as long as it is not preceded by 'p' or 'r'
>>> re.sub(r'(?<![pr]).', '*', 'spare')
'**a*e'

>>> re.sub(r'.(?<![pr].)', '*', 'spare')
'**a*e'

# replace 'par' as long as 's' is not present later in the input
# this assumes that the lookahead doesn't conflict with search pattern
# i.e. 's' will not conflict 'par' but would affect if it was 'r' and 'par'
>>> re.sub(r'par(?!.s)', 'X', 'par spare part party')
'par sXe Xt Xty'

>>> re.sub(r'(?!.s)par', 'X', 'par spare part party')
'par sXe Xt Xty'

# since the three assertions used here are all zero-width,
# all of the 6 possible combinations will be equivalent
>>> re.sub(r'(?!\Z)\b(?<!\A)', ' ', 'foo_baz=num1+35*42/num2')
'foo_baz = num1 + 35 * 42 / num2'
```

Positive lookarounds

Unlike negative lookarounds, absence of something will not satisfy positive lookarounds. Instead, for the condition to satisfy, the pattern has to match actual characters and/or zero-width assertions. Positive lookahead can be identified by use of `=` in the grouping. The complete syntax looks like:

- `(?=pat)` for positive lookahead assertion
- `(?<=pat)` for positive lookbehind assertion

```
# extract digits only if it is followed by ,
# note that end of string doesn't qualify as this is positive assertion
>>> re.findall(r'\d+(?=,)', '42 foo-5, baz3; x-83, y-20: f12')
['5', '83']

# extract digits only if it is preceded by - and followed by ; or :
>>> re.findall(r'(?<=-)\d+(?=[:;])', '42 foo-5, baz3; x-83, y-20: f12')
['20']

# replace 'par' as long as 'part' occurs as a whole word later in the line
>>> re.sub(r'par(?=.*\bpart\b)', 'X', 'par spare part party')
'X sXe part party'
```

Lookarounds can be quite handy in field based processing.

```
# except first and last fields
>>> re.findall(r'(?<=,)[^,]+(?=,)', '1,two,3,four,5')
['two', '3', 'four']

# replace empty fields with NA
# note that in this case, order of lookbehind and lookahead doesn't matter
>>> re.sub(r'(?<![^,])(?![,])', 'NA', ',1,,,two,3,,,')
'NA,1,NA,NA,two,3,NA,NA,NA'
```

Capture groups inside positive lookarounds

Even though lookarounds are not part of matched portions, capture groups can be used inside positive lookarounds. Can you reason out why it won't work for negative lookarounds?

```
>>> print(re.sub(r'(\S+\s+)(?=(\S+)\s)', r'\1\2\n', 'a b c d e'))
a b
b c
c d
d e

# and of course, use non-capturing group where needed
>>> re.findall(r'(?<=(po|ca)re)\d+', 'pore42 car3 pare7 care5')
['po', 'ca']
>>> re.findall(r'(?<=(?:po|ca)re)\d+', 'pore42 car3 pare7 care5')
['42', '5']
```

Conditional AND with lookarounds

As promised earlier, here's how lookarounds make it simpler to construct AND conditionals.

```
>>> words = ['sequoia', 'subtle', 'questionable', 'exhibit', 'equation']

# words containing 'b' and 'e' and 't' in any order
# same as: r'b.*e.*t|b.*t.*e|e.*b.*t|e.*t.*b|t.*b.*e|t.*e.*b'
>>> [w for w in words if re.search(r'(?=.*b)(?=.*e).*t', w)]
['subtle', 'questionable', 'exhibit']

# words containing all lowercase vowels in any order
>>> [w for w in words if re.search(r'(?=.*a)(?=.*e)(?=.*i)(?=.*o).*u', w)]
['sequoia', 'questionable', 'equation']

# words containing 'a' and 'q' but not 'n' at the end of the element
>>> [w for w in words if re.search(r'(?=.*a)(?=.*q)(?!.*n$)', w)]
['sequoia', 'questionable']
```

Variable length lookbehind

With lookbehind assertion (both positive and negative), the pattern used for the assertion cannot *imply* matching variable length of text. Fixed length quantifier is allowed. Different length alternations are not allowed, even if the individual alternations are of fixed length. Here's some examples to clarify these points.

```
>>> s = 'pore42 tar3 dare7 care5'

# allowed
>>> re.findall(r'(?<=(?:po|da)re)\d+', s)
['42', '7']
>>> re.findall(r'(?<=\b[a-z]{4})\d+', s)
['42', '7', '5']

# not allowed
>>> re.findall(r'(?<!\tar|dare)\d+', s)
re.error: look-behind requires fixed-width pattern
>>> re.findall(r'(?<=\b[pd][a-z]*)\d+', s)
re.error: look-behind requires fixed-width pattern
>>> re.sub(r'(?<=\A|,)(?=[\Z])', 'NA', ',1,,,two,3,,,')
re.error: look-behind requires fixed-width pattern
```

There are various workarounds possible depending upon the use case. See [Variable length lookbehind](#) section for `regex` module for the easiest and most robust solution.

First up, when the lookbehind alternatives are of fixed length. In such cases, you can manually expand the alternatives into individual lookbehinds. For negatives, place them next to each other. For positives, separate them with alternation inside another grouping.

```
>>> s = 'pore42 tar3 dare7 care5'

# workaround for r'(?<!tar|dare)\d+'
>>> re.findall(r'(?<!tar)(?<!dare)\d+', s)
['42', '5']

# workaround for r'(?<=tar|dare)\d+'
>>> re.findall(r'(?:(?<=tar)|(?<=dare))\d+', s)
['3', '7']

# workaround for r'(?<=\A|,)(?<=,|\Z)'
>>> re.sub(r'((?<=\A)|(?<=,))(?<=,|\Z)', 'NA', ',1,,,two,3,,,')
'NA,1,NA,NA,two,3,NA,NA,NA'
```

For some of the positive lookbehind cases, you can skip lookbehind entirely and workaround with normal groupings. This works even when you don't know the length of patterns.

```
>>> s = 'pore42 tar3 dare7 care5'

# same as: re.findall(r'(?:(?<=tar)|(?<=dare))\d+', s)
>>> re.findall(r'(?:(?<=tar)|(?<=dare))\d+', s)
['3', '7']

# delete digits only if they are preceded by 'tar' or 'dare'
>>> re.sub(r'(tar|dare)\d+', r'\1', s)
'pore42 tar dare care5'

# workaround for r'(?<=\b[dp][a-z]*)\d+'
# get digits only if they are preceded by a word starting with 'p' or 'd'
>>> re.findall(r'\b[dp][a-z]*(\d+)', s)
['42', '7']

# delete digits only if they are preceded by a word starting with 'p' or 'd'
>>> re.sub(r'(\b[dp][a-z]*)\d+', r'\1', s)
'pore tar3 dare care5'
```

However, if you don't know the lengths for negative lookbehind, you cannot use the above workarounds. The next section will show how to negate a grouping, and that helps for some of the variable negative lookbehind cases.

Negated groups

Variable length negative lookbehind can be simulated using negative lookahead (which doesn't have restriction on variable length) inside a grouping and applying quantifier to match characters one by one. This also showcases how grouping can be negated in certain cases. Note that this will only work if you have well defined conditions before the negated group.

```
# note the use of \A anchor to force matching all characters up to 'dog'
>>> bool(re.search(r'\A((?!cat).)*dog', 'fox,cat,dog,parrot'))
False
```

```
>>> bool(re.search(r'\A(?:!parrot.)*dog', 'fox,cat,dog,parrot'))
True
# without the anchor, you'll get false matches
>>> bool(re.search(r'((?!cat.)*dog', 'fox,cat,dog,parrot'))
True

# easier to understand by checking matched portion
>>> re.search(r'\A(?:!cat.)*', 'fox,cat,dog,parrot')[0]
'fox,'
>>> re.search(r'\A(?:!parrot.)*', 'fox,cat,dog,parrot')[0]
'fox,cat,dog,'
>>> re.search(r'\A(?:!(.)\2.)*', 'fox,cat,dog,parrot')[0]
'fox,cat,dog,pa'
```

Here's some examples of matching two patterns with a negated group between them.

```
# match if 'do' is not there between 'at' and 'par'
>>> bool(re.search(r'at(?:!do.)*par', 'fox,cat,dog,parrot'))
False

# match if 'go' is not there between 'at' and 'par'
>>> bool(re.search(r'at(?:!go.)*par', 'fox,cat,dog,parrot'))
True
>>> re.search(r'at(?:!go.)*par', 'fox,cat,dog,parrot')[0]
'at,dog,par'

# use non-capturing group if required
>>> re.findall(r'a(?:?!d.)*z', 'at,baz,a2z,bad-zoo')
['at,baz', 'ad-z']
```

Cheatsheet and Summary

Note	Description
lookarounds	custom assertions, zero-width like anchors
(?!pat)	negative lookahead assertion
(?<!pat)	negative lookbehind assertion
(?=pat)	positive lookahead assertion
(?<=pat)	positive lookbehind assertion
(?!pat1)(?=pat2)	multiple assertions can be specified next to each other in any order as they mark a matching location without consuming characters
((?!pat).)*	Negate a grouping, similar to negated character class

In this chapter, you learnt how to use lookarounds to create custom restrictions and also how to use negated grouping. With this, most of the powerful features of regular expressions have been covered. The special groupings seem never ending though, there's some more of them in coming chapters!!

Exercises



Please use lookarounds for solving the following exercises even if you can do it without lookarounds. Unless you cannot use lookarounds for cases like variable length lookbehinds.

a) Replace all whole words with `X` unless it is preceded by `(` character.

```
>>> ip = '(apple) guava berry) apple (mango) (grape'

##### add your solution here
'(apple) X X) X (mango) (grape'
```

b) Replace all whole words with `X` unless it is followed by `)` character.

```
>>> ip = '(apple) guava berry) apple (mango) (grape'

##### add your solution here
'(apple) X berry) X (mango) (X'
```

c) Replace all whole words with `X` unless it is preceded by `(` or followed by `)` characters.

```
>>> ip = '(apple) guava berry) apple (mango) (grape'

##### add your solution here
'(apple) X berry) X (mango) (grape'
```

d) Extract all whole words that do not end with `e` or `n`.

```
>>> ip = 'at row on urn e note dust n'

##### add your solution here
['at', 'row', 'dust']
```

e) Extract all whole words that do not start with `a` or `d` or `n`.

```
>>> ip = 'at row on urn e note dust n'

##### add your solution here
['row', 'on', 'urn', 'e']
```

f) Extract all whole words only if they are followed by `:` or `,` or `-`.

```
>>> ip = 'poke,on=-so:ink.to/is(vast)ever-sit'

##### add your solution here
['poke', 'so', 'ever']
```

g) Extract all whole words only if they are preceded by `=` or `/` or `-`.

```
>>> ip = 'poke,on=-so:ink.to/is(vast)ever-sit'

##### add your solution here
['so', 'is', 'sit']
```

h) Extract all whole words only if they are preceded by `=` or `:` and followed by `:` or `.`.

```
>>> ip = 'poke,on==so:ink.to/is(vast)ever-sit'

##### add your solution here
['so', 'ink']
```

i) Extract all whole words only if they are preceded by `=` or `:` or `.` or `(` or `-` and not followed by `.` or `/`.

```
>>> ip = 'poke,on==so:ink.to/is(vast)ever-sit'

##### add your solution here
['so', 'vast', 'sit']
```

j) Remove leading and trailing whitespaces from all the individual fields where `,` is the field separator.

```
>>> csv1 = ' comma ,separated ,values \t\r '
>>> csv2 = 'good bad,nice ice , 42 , , stall small'

>>> remove_whitespace = re.compile() ##### add your solution here

>>> remove_whitespace.sub('', csv1)
'comma,separated,values'
>>> remove_whitespace.sub('', csv2)
'good bad,nice ice,42,,stall small'
```

k) Filter all elements that satisfy all of these rules:

- should have at least two alphabets
- should have at least 3 digits
- should have at least one special character among `%` or `*` or `#` or `$`
- should not end with a whitespace character

```
>>> pwds = ['hunter2', 'F2H3u%9', '*X3Yz3.14\t', 'r2_d2_42', 'A $B C1234']

##### add your solution here
['F2H3u%9', 'A $B C1234']
```

l) For the given string, surround all whole words with `{}` except for whole words `par` and `cat` and `apple`.

```
>>> ip = 'part; cat {super} rest_42 par scatter apple spar'

##### add your solution here
'{part}; cat {{super}} {rest_42} par {scatter} apple {spar}'
```

m) Extract integer portion of floating-point numbers for the given string. A number ending with `.` and no further digits should not be considered.

```
>>> ip = '12 ab32.4 go 5 2. 46.42 5'

##### add your solution here
['32', '46']
```

n) For the given input strings, extract all overlapping two character sequences.

```
>>> s1 = 'apple'
>>> s2 = '1.2-3:4'

>>> pat = re.compile()          ##### add your solution here

##### add your solution here for s1
['ap', 'pp', 'pl', 'le']
##### add your solution here for s2
['1.', '.2', '2-', '-3', '3:', ':4']
```

o) The given input strings contain fields separated by `:` character. Delete `:` and the last field if there is a digit character anywhere before the last field.

```
>>> s1 = '42:cat'
>>> s2 = 'twelve:a2b'
>>> s3 = 'we:be:he:0:a:b:bother'

>>> pat = re.compile()          ##### add your solution here

>>> pat.sub()                    ##### add your solution here for s1
'42'
>>> pat.sub()                    ##### add your solution here for s2
'twelve:a2b'
>>> pat.sub()                    ##### add your solution here for s3
'we:be:he:0:a:b'
```

p) Extract all whole words unless they are preceded by `:` or `<=>` or `----` or `#` .

```
>>> ip = '::.very--at<=>row|in.a_b#b2c=>lion----east'

##### add your solution here
['at', 'in', 'a_b', 'lion']
```

q) Match strings if it contains `qty` followed by `price` but not if there is **whitespace** or the string `error` between them.

```
>>> str1 = '23,qty,price,42'
>>> str2 = 'qty price,oh'
>>> str3 = '3.14,qty,6,errors,9,price,3'
>>> str4 = '42\nqty-6,apple-56,price-234,error'
>>> str5 = '4,price,3.14,qty,4'

>>> neg = re.compile()          ##### add your solution here

>>> bool(neg.search(str1))
True
>>> bool(neg.search(str2))
False
>>> bool(neg.search(str3))
False
>>> bool(neg.search(str4))
```

True

```
>>> bool(neg.search(str5))
```

False

r) Can you reason out why the output shown is different for these two regular expressions?

```
>>> ip = 'I have 12, he has 2!'
```

```
>>> re.sub(r'\b..\b', '{\g<0>}', ip)
```

```
'{I }have {12}{, }{he} has{ 2}!'
```

```
>>> re.sub(r'(?!\w)..\w', '{\g<0>}', ip)
```

```
'I have {12}, {he} has {2!}'
```

Flags

Just like options change the default behavior of commands used from a terminal, flags are used to change aspects of RE behavior. You have already seen flags for ignoring case and changing behavior of line anchors. Flags can be applied to entire RE using `flags` optional argument or to a particular portion of RE using special groups. And both of these forms can be mixed up as well. In regular expression parlance, flags are also known as **modifiers**.

Flags already seen will again be discussed in this chapter for sake of completeness. You'll also see how to combine multiple flags.

re.IGNORECASE

First up, the flag to ignore case while matching alphabets. When `flags` argument is used, this can be specified as `re.I` or `re.IGNORECASE` constants.

```
>>> bool(re.search(r'cat', 'Cat'))
False
>>> bool(re.search(r'cat', 'Cat', flags=re.IGNORECASE))
True

>>> re.findall(r'c.t', 'Cat cot CATER ScUtTle', flags=re.I)
['Cat', 'cot', 'CAT', 'cUt']

# without flag, you need to use: r'[a-zA-Z]+'
# with flag, can also use: r'[A-Z]+'
>>> re.findall(r'[a-z]+', 'Sample123string42with777numbers', flags=re.I)
['Sample', 'string', 'with', 'numbers']
```

re.DOTALL

Use `re.S` or `re.DOTALL` to allow `.` metacharacter to match newline character as well.

```
# by default, the . metacharacter doesn't match newline
>>> re.sub(r'the.*ice', 'X', 'Hi there\nHave a Nice Day')
'Hi there\nHave a Nice Day'

# re.S flag will allow newline character to be matched as well
>>> re.sub(r'the.*ice', 'X', 'Hi there\nHave a Nice Day', flags=re.S)
'Hi X Day'
```

Multiple flags can be combined using bitwise OR operator.

```
>>> re.sub(r'the.*day', 'Bye', 'Hi there\nHave a Nice Day', flags=re.S|re.I)
'Hi Bye'
```

re.MULTILINE

As seen earlier, `re.M` or `re.MULTILINE` flag would allow `^` and `$` anchors to match line wise instead of whole string.

```
# check if any line in the string starts with 'top'
>>> bool(re.search(r'^top', "hi hello\ntop spot", flags=re.M))
True

# check if any line in the string ends with 'ar'
>>> bool(re.search(r'ar$', "spare\npar\ndare", flags=re.M))
True
```

re.VERBOSE

The `re.X` or `re.VERBOSE` flag is another provision like the named capture groups to help add clarity to RE definitions. This flag allows to use literal whitespaces for aligning purposes and add comments after the `#` character to break down complex RE into multiple lines.

```
# same as: pat = re.compile(r'\A(?:[^\,]+,){3}([^\,]+)')
# note the use of triple quoted string
>>> pat = re.compile(r'''
...     \A(                # group-1, captures first 3 columns
...        (?:[^\,]+,){3}  # non-capturing group to get the 3 columns
...     )
...     ([^\,]+)          # group-2, captures 4th column
...     ''', flags=re.X)

>>> pat.sub(r'\1(\2)', '1,2,3,4,5,6,7')
'1,2,3,(4),5,6,7'
```

For precise definition, here's the relevant quote from documentation:

Whitespace within the pattern is ignored, except when in a character class, or when preceded by an unescaped backslash, or within tokens like `*?`, `(?:` or `(?P<...>`. When a line contains a `#` that is not in a character class and is not preceded by an unescaped backslash, all characters from the leftmost such `#` through the end of the line are ignored.

```
>>> bool(re.search(r't a', 'cat and dog', flags=re.X))
False
>>> bool(re.search(r't\ a', 'cat and dog', flags=re.X))
True
>>> bool(re.search(r't[ ]a', 'cat and dog', flags=re.X))
True
>>> bool(re.search(r't\x20a', 'cat and dog', flags=re.X))
True

>>> re.search(r'a#b', 'foo a#b 123', flags=re.X)[0]
'a'
```

```
>>> re.search(r'a#b', 'foo a#b 123', flags=re.X)[0]
'a#b'
```

Inline comments

Comments can also be added using `(?#comment)` special group. This is independent of the `re.X` flag.

```
>>> pat = re.compile(r'\A(?:[^\s,]+){3}(?#3-cols)([^\s,]+)(?#4th-col)')
>>> pat.sub(r'\1(\2)', '1,2,3,4,5,6,7')
'1,2,3,(4),5,6,7'
```

Inline flags

To apply flags to specific portions of RE, specify them inside a special grouping syntax. This will override the flags applied to entire RE definitions, if any. The syntax variations are:

- `(?flags:pat)` will apply flags only for this portion
- `(?-flags:pat)` will negate flags only for this portion
- `(?flags-flags:pat)` will apply and negate particular flags only for this portion
- `(?flags)` will apply flags for whole RE definition, can only be specified at the start of RE definition
 - if anchors are needed, they should be specified after `(?flags)`

In these ways, flags can be specified precisely only where it is needed. The flags are to be given as single letter lowercase version of short form constants. For example, `i` for `re.I`, `s` for `re.S` and so on, except `L` for `re.L` or `re.LOCALE` (discussed in [re.ASCII](#) section). And as can be observed from below examples, these do not act as capture groups.

```
# case-sensitive for whole RE definition
>>> re.findall(r'Cat[a-z]*\b', 'Cat SCatTeR CATER cAts')
['Cat']
# case-insensitive only for '[a-z]*' portion
>>> re.findall(r'Cat(?i:[a-z]*)\b', 'Cat SCatTeR CATER cAts')
['Cat', 'CatTeR']
# case-insensitive for whole RE definition using flags argument
>>> re.findall(r'Cat[a-z]*\b', 'Cat SCatTeR CATER cAts', flags=re.I)
['Cat', 'CatTeR', 'CATER', 'cAts']
# case-insensitive for whole RE definition using inline flags
>>> re.findall(r'(?i)Cat[a-z]*\b', 'Cat SCatTeR CATER cAts')
['Cat', 'CatTeR', 'CATER', 'cAts']
# case-sensitive only for 'Cat' portion
>>> re.findall(r'(?-i:Cat)[a-z]*\b', 'Cat SCatTeR CATER cAts', flags=re.I)
['Cat', 'CatTeR']
```

Cheatsheet and Summary

Note	Description
<code>re.IGNORECASE</code> or <code>re.I</code>	flag to ignore case
<code>re.DOTALL</code> or <code>re.S</code>	allow <code>.</code> metacharacter to match newline character
<code>flags=re.S re.I</code>	multiple flags can be combined using <code> </code> operator
<code>re.MULTILINE</code> or <code>re.M</code>	allow <code>^</code> and <code>\$</code> anchors to match line wise
<code>re.VERBOSE</code> or <code>re.X</code>	allows to use literal whitespaces for aligning purposes and to add comments after the <code>#</code> character escape spaces and <code>#</code> if needed as part of actual RE
<code>(?#comment)</code>	another way to add comments, not a flag
<code>(?flags:pat)</code>	inline flags only for this <code>pat</code> , overrides <code>flags</code> argument where flags is <code>i</code> for <code>re.I</code> , <code>s</code> for <code>re.S</code> , etc except <code>L</code> for <code>re.L</code>
<code>(?-flags:pat)</code>	negate flags only for this <code>pat</code>
<code>(?flags-flags:pat)</code>	apply and negate particular flags only for this <code>pat</code>
<code>(?flags)</code>	apply flags for whole RE, can be used only at start of RE anchors if any, should be specified after <code>(?flags)</code>

This chapter showed some of the flags that can be used to change default behavior of RE definition. And more special groupings were covered.

Exercises

a) Remove from first occurrence of `hat` to last occurrence of `ice` for the given input strings. Match these markers case insensitively.

```
>>> s1 = 'But Cool THAT\nsee What okay\nwow quite'
>>> s2 = 'it this hat is sliced HIT.'

>>> pat = re.compile()          ##### add your solution here

>>> pat.sub('', s1)
'But Cool Te'
>>> pat.sub('', s2)
'it this .'
```

b) Delete from `start` if it is at the beginning of a line up to the next occurrence of the `end` at the end of a line. Match these markers case insensitively.

```
>>> para = '''\
... good start
... start working on that
... project you always wanted
... to, do not let it end
... hi there
... start and end the end
... 42
```

```

... Start and try to
... finish the End
... bye'''

>>> pat = re.compile()          ##### add your solution here

>>> print(pat.sub(' ', para))
good start

hi there

42

bye

```

c) For the given input strings, match all of these three patterns:

- This case sensitively
- nice and cool case insensitively

```

>>> s1 = 'This is nice and Cool'
>>> s2 = 'Nice and cool this is'
>>> s3 = 'What is so nice and cool about This?'

>>> pat = re.compile()          ##### add your solution here

>>> bool(pat.search(s1))
True
>>> bool(pat.search(s2))
False
>>> bool(pat.search(s3))
True

```

d) For the given input strings, match if the string begins with Th and also contains a line that starts with There .

```

>>> s1 = 'There there\nHave a cookie'
>>> s2 = 'This is a mess\nYeah?\nThereeeee'
>>> s3 = '0h\nThere goes the fun'

>>> pat = re.compile()          ##### add your solution here

>>> bool(pat.search(s1))
True
>>> bool(pat.search(s2))
True
>>> bool(pat.search(s3))
False

```

e) Explore what the re.DEBUG flag does. Here's some example patterns to check out.

- re.compile(r'\Aden|ly\Z', flags=re.DEBUG)

- `re.compile(r'\b(0x)?[\da-f]+\b', flags=re.DEBUG)`
- `re.compile(r'\b(?:0x)?[\da-f]+\b', flags=re.I|re.DEBUG)`

Unicode

So far in the book, all examples were meant for strings made up of ASCII characters only. However, `re` module matching is Unicode by default. See [docs.python: Unicode](#) for a tutorial on Unicode support in Python. This chapter will briefly discuss a few things related to Unicode matching.

re.ASCII

Flags can be used to override the default Unicode setting. The `re.A` or `re.ASCII` flag will change `\b`, `\w`, `\d`, `\s` and their opposites to match only based on ASCII characters.

```
# \w is Unicode aware
>>> re.findall(r'\w+', 'fox:αλεπού')
['fox', 'αλεπού']

# restrict matching to only ASCII characters
>>> re.findall(r'\w+', 'fox:αλεπού', flags=re.A)
['fox']

# or, explicitly define the characters to match using character class
>>> re.findall(r'[a-zA-Z0-9_]+', 'fox:αλεπού')
['fox']
```

However, the four characters shown in the code snippet below are also matched when `re.I` is used without `re.A` flag. The relevant quote from the docs is:

Note that when the Unicode patterns `[a-z]` or `[A-Z]` are used in combination with the IGNORECASE flag, they will match the 52 ASCII letters and 4 additional non-ASCII letters: `İ` (U+0130, Latin capital letter I with dot above), `ı` (U+0131, Latin small letter dotless i), `ſ` (U+017F, Latin small letter long s) and `℄` (U+212A, Kelvin sign). If the ASCII flag is used, only letters `a` to `z` and `A` to `Z` are matched.

```
>>> bool(re.search(r'[a-zA-Z]', 'İıſ℄'))
False

>>> re.search(r'[a-z]+', 'İıſ℄', flags=re.I)[0]
'İıſ℄'

>>> bool(re.search(r'[a-z]', 'İıſ℄', flags=re.I|re.A))
False
```



Use `re.L` or `re.LOCALE` to work based on locale settings for bytes data type.

Codepoints and Unicode escapes

You can use escapes `\u` and `\U` to specify Unicode characters with 4 and 8 hexadecimal digits respectively. The below snippet also shows how to get codepoints (numerical value of a character) in Python.

```
# to get codepoints for ASCII characters
>>> [ord(c) for c in 'fox']
[102, 111, 120]
>>> [hex(ord(c)) for c in 'fox']
['0x66', '0x6f', '0x78']

# to get codepoints for Unicode characters
>>> [c.encode('unicode_escape') for c in 'αλεπού']
[b'\\u03b1', b'\\u03b2', b'\\u03b5', b'\\u03c0', b'\\u03b2', b'\\u03c4']
>>> [c.encode('unicode_escape') for c in 'İıfK']
[b'\\u0130', b'\\u0131', b'\\u017f', b'\\u212a']

# character range example using \u
# all english lowercase letters
>>> re.findall(r'[\u0061-\u007a]+', 'fox:αλεπού,eagle:αετός')
['fox', 'eagle']
```



See also: [codepoints](#), a site dedicated for Unicode characters.

\N escape sequence

You can also specify a Unicode character using `\N{name}` escape sequence. See [unicode: NamesList](#) for a full list of names. From the docs:

Changed in version 3.8: The `\N{name}` escape sequence has been added. As in string literals, it expands to the named Unicode character (e.g. `\N{EM DASH}`).

```
# can also use '\N{em dash}'
>>> '\N{EM DASH}'
'_'

>>> '\N{LATIN SMALL LETTER TURNED DELTA}'
'g'
```

Cheatsheet and Summary

Note	Description
docs.python: Unicode <code>re.ASCII</code> or <code>re.A</code>	tutorial on Unicode support in Python match only ASCII characters for <code>\b</code> , <code>\w</code> , <code>\d</code> , <code>\s</code>

Note	Description
<code>re.LOCALE</code> or <code>re.L</code>	and their opposites, when using Unicode patterns use locale settings for byte patterns and 8-bit locales
<code>ifK</code>	characters that can match if <code>re.I</code> is used but not <code>re.A</code>
<code>ord(c)</code>	get codepoint for ASCII character <code>c</code>
<code>c.encode('unicode_escape')</code>	get codepoint for Unicode character <code>c</code>
<code>\uXXXX</code>	codepoint defined using 4 hexadecimal digits
<code>\UXXXXXXXX</code>	codepoint defined using 8 hexadecimal digits
<code>\N{name}</code>	Unicode character defined by its name
	See unicode: NamesList for full list

A comprehensive discussion on RE usage with Unicode characters is out of scope for this book. Resources like [regular-expressions: unicode](#) and [Programmers introduction to Unicode](#) are recommended for further study. See also [Unicode character sets](#) section of `regex` module.

Exercises

a) Output `True` or `False` depending on input string made up of ASCII characters or not. Consider the input to be non-empty strings and any character that isn't part of 7-bit ASCII set should give `False`. Do you need regular expressions for this?

```
>>> str1 = '123-456'
>>> str2 = 'good food'
>>> str3 = 'happy learning!'
>>> str4 = 'ïfK'

##### add your solution here for str1
False
##### add your solution here for str2
False
##### add your solution here for str3
True
##### add your solution here for str4
False
```

b) Does `.` quantifier with `re.ASCII` flag enabled match non-ASCII characters?

c) Explore the following Q&A threads.

- [stackoverflow: remove powered number from string](#)
- [stackoverflow: regular expression for French characters](#)

regex module

The third party `regex` module (<https://pypi.org/project/regex/>) offers advanced features like those found in Perl language and other regular expression implementations. To install the module from command line, you can use either of these depending on your usage:

- `pip install regex` in a virtual environment
- `python3.8 -m pip install --user regex` for system wide accessibility

By default, `regex` module uses `VERSION0` which is compatible with the `re` module. If you want all the features, `VERSION1` should be used. For example, set operators is a feature available only with `VERSION1`. You can choose the version to be used in two ways. Setting `regex.DEFAULT_VERSION` to `regex.VERSION0` or `regex.VERSION1` is a global option. `(?V0)` and `(?V1)` are inline flag options.



The examples in this chapter are presented assuming `VERSION1` is enabled.

```
>>> import regex
>>> regex.DEFAULT_VERSION = regex.VERSION1

>>> sentence = 'This is a sample string'
>>> bool(regex.search(r'is', sentence))
True
```

Possessive quantifiers

Appending a `+` to greedy quantifiers makes them possessive. These behave like greedy quantifiers, but without the backtracking. So, something like `r'Error.*+valid'` will never match because `.*` will consume all the remaining characters. If both greedy and possessive quantifier versions are functionally equivalent, then possessive is preferred because it will fail faster for non-matching cases.

```
# functionally equivalent greedy and possessive versions
>>> demo = ['abc', 'ac', 'adc', 'abbc', 'xabbbcz', 'bbb', 'bc', 'abbbbbc']
>>> [w for w in demo if regex.search(r'ab*c', w)]
['abc', 'ac', 'abbc', 'xabbbcz', 'abbbbbc']
>>> [w for w in demo if regex.search(r'ab*c+', w)]
['abc', 'ac', 'abbc', 'xabbbcz', 'abbbbbc']

# different results
# numbers >= 100 if there are leading zeros
>>> regex.findall(r'\b0\d{3,}\b', '0501 035 154 12 26 98234')
['0501', '035', '154', '98234']
>>> regex.findall(r'\b0*+\d{3,}\b', '0501 035 154 12 26 98234')
['0501', '154', '98234']
```

The effect of possessive quantifier can also be expressed using **atomic grouping**. The syntax is `(?>pat)`, where `pat` is the portion you want to match possessively.

```
# same as: r'[bo]++'
>>> regex.sub(r'(>[bo]+)', 'X', 'abbbc fooooooot')
'aXc fXt'

# same as: r'\b0*\d{3,}\b'
>>> regex.findall(r'\b(>0*)\d{3,}\b', '0501 035 154 12 26 98234')
['0501', '154', '98234']
```

Subexpression calls

If backreferences are like variables, then subexpression calls are like functions. Backreferences allow you to reuse the portion matched by the capture group. Subexpression calls allow you to reuse the pattern that was used inside the capture group. You can call subexpressions recursively too, see [Recursive matching](#) section for details.

The syntax is `(?N)` where `N` is the capture group you want to call. This is applicable only in RE definition, not in replacement sections.

```
>>> row = 'today,2008-03-24,food,2012-08-12,nice,5632'

# with re module and manually repeating the pattern
>>> re.search(r'\d{4}-\d{2}-\d{2}.*\d{4}-\d{2}-\d{2}', row)[0]
'2008-03-24,food,2012-08-12'

# with regex module and subexpression calling
>>> regex.search(r'(\d{4}-\d{2}-\d{2}).*(?1)', row)[0]
'2008-03-24,food,2012-08-12'
```

Named capture groups can be called using `(?&name)` syntax.

```
>>> row = 'today,2008-03-24,food,2012-08-12,nice,5632'

>>> regex.search(r'(?P<date>\d{4}-\d{2}-\d{2}).*(?&date)', row)[0]
'2008-03-24,food,2012-08-12'
```

Positive lookbehind with \K

Most (but not all) of the positive lookbehind cases can be solved by adding `\K` as a suffix to the pattern to be tested. This will work for variable length patterns as well.

```
# similar to: r'(?<=\b\w)\w*\W*'
# text matched before \K won't be replaced
>>> regex.sub(r'\b\w\K\w*\W*', '', 'sea eat car rat eel tea')
'secret'

# variable length example
# replace only 3rd occurrence of 'cat'
>>> regex.sub(r'(cat.*?){2}\Kcat', 'X', 'cat scatter cater scat', count=1)
'cat scatter Xer scat'
```

Here's another example that won't work if greedy quantifier is used instead of possessive quantifier.

```
>>> row = '421,foo,2425,42,5,foo,6,6,42'

# lookarounds used to ensure start/end of column matching
# possessive quantifier used to ensure partial column is not captured
# if a column has same text as another column, the latter column is deleted
>>> while True:
...     row, cnt = regex.subn(r'(?![^,])([^\,]+).*\K,\1(?![^,])', r'', row)
...     if cnt == 0:
...         break
...
>>> row
'421,foo,2425,42,5,6'
```

Variable length lookbehind

The `regex` module allows using variable length lookbehind without needing any change.

```
>>> s = 'pore42 tar3 dare7 care5'
>>> regex.findall(r'(?<!(tar|dare)\d+', s)
['42', '5']
>>> regex.findall(r'(?<=\\b[\\p{\\d}][a-z]*)\\d+', s)
['42', '7']
>>> regex.sub(r'(?<=\\A|,)(?=,|\\Z)', 'NA', ',1,,,two,3,,,')
'NA,1,NA,NA,two,3,NA,NA,NA'

>>> regex.sub(r'(?<=(cat.*?){2})cat', 'X', 'cat scatter cater scat', count=1)
'cat scatter Xer scat'

>>> bool(regex.search(r'(?<!(cat.*)dog', 'fox,cat,dog,parrot'))
False
>>> bool(regex.search(r'(?<!(parrot.*)dog', 'fox,cat,dog,parrot'))
True
```



As lookarounds do not consume characters, don't use variable length lookbehind between two patterns. Use negated groups instead.

```
# match if 'go' is not there between 'at' and 'par'

# wrong use of lookahead
>>> bool(regex.search(r'at(?<!(go.*)par', 'fox,cat,dog,parrot'))
False

# correct use of negated group
>>> bool(regex.search(r'at((?!go).)*par', 'fox,cat,dog,parrot'))
True
```

\G anchor

The `\G` anchor restricts matching from start of string like the `\A` anchor. In addition, after a match is done, ending of that match is considered as the new anchor location. This process is repeated again and continues until the given RE fails to match (assuming multiple matches with `sub`, `findall` etc).

```
# all non-whitespace characters from start of string
>>> regex.findall(r'\G\S', '123-87-593 42 foo')
['1', '2', '3', '-', '8', '7', '-', '5', '9', '3']
>>> regex.sub(r'\G\S', '*', '123-87-593 42 foo')
'***** 42 foo'

# all digits and optional hyphen combo from start of string
>>> regex.findall(r'\G\d+?', '123-87-593 42 foo')
['123-', '87-', '593']
>>> regex.sub(r'\G(\d+)(-?)', r'(\1)\2', '123-87-593 42 foo')
'(123)-(87)-(593) 42 foo'

# all word characters from start of string
# only if it is followed by word character
>>> regex.findall(r'\G\w(?:\w)', 'cat12 bat pin')
['c', 'a', 't', '1']
>>> regex.sub(r'\G\w(?:\w)', r'\g<0>:', 'cat12 bat pin')
'c:a:t:1:2 bat pin'

# all lowercase alphabets or space from start of string
>>> regex.sub(r'\G[a-z ]', r'(\g<0>)', 'par tar-den hen-food mood')
'(p)(a)(r)( ) (t)(a)(r)-den hen-food mood'
```

Recursive matching

The subexpression call special group was introduced as analogous to function call. And in typical function fashion, it does support recursion. Useful to match nested patterns, which is usually not recommended to be done with regular expressions. Indeed, use a proper parser library if you are looking to parse file formats like html, xml, json, csv, etc. But for some cases, a parser might not be available and using RE might be simpler than writing a parser from scratch.

First up, a RE to match a set of parentheses that is not nested (termed as **level-one** RE for reference).

```
# note the use of possessive quantifier
>>> eqn0 = 'a + (b * c) - (d / e)'
>>> regex.findall(r'\([^()]++\)', eqn0)
['(b * c)', '(d / e)']

>>> eqn1 = '((f+x)^y-42)*((3-g)^z+2)'
>>> regex.findall(r'\([^()]++\)', eqn1)
['(f+x)', '(3-g)']
```

Next, matching a set of parentheses which may optionally contain any number of non-nested sets of parentheses (termed as **level-two** RE for reference). See [debuggex](#) for a railroad diagram, which visually shows the recursive nature of this RE.

```
>>> eqn1 = '((f+x)^y-42)*((3-g)^z+2)'
# note the use of non-capturing group
>>> regex.findall(r'\((?:[^\(\)]++|\([^()]+\))+\)', eqn1)
['((f+x)^y-42)', '((3-g)^z+2)']

>>> eqn2 = 'a + (b) + ((c)) + (((d)))'
>>> regex.findall(r'\((?:[^\(\)]++|\([^()]+\))+\)', eqn2)
['(b)', '((c))', '(((d)))']
```

That looks very cryptic. Better to use `regex.X` flag for clarity as well as for comparing against the recursive version. Breaking down the RE, you can see `(` and `)` have to be matched literally. Inside that, valid string is made up of either non-parentheses characters or a non-nested parentheses sequence (level-one RE).

```
>>> lvl2 = regex.compile('''
...     \(\s*          #literal (
...     (?:           #start of non-capturing group
...         [^\(\)]++  #non-parentheses characters
...         |          #OR
...         \([^()]+\) #level-one RE
...     )++          #end of non-capturing group, 1 or more times
...     \)           #literal )
... ''', flags=regex.X)

>>> lvl2.findall(eqn1)
['((f+x)^y-42)', '((3-g)^z+2)']

>>> lvl2.findall(eqn2)
['(b)', '((c))', '(((d)))']
```

To recursively match any number of nested sets of parentheses, use a capture group and call it within the capture group itself. Since entire RE needs to be called here, you can use the default zeroth capture group (this also helps to avoid having to use `finditer`). Comparing with level-two RE, the only change is that `(?0)` is used instead of the level-one RE in the second alternation.

```
>>> lvlN = regex.compile('''
...     \(\s*          #literal (
...     (?:           #start of non-capturing group
...         [^\(\)]++  #non-parentheses characters
...         |          #OR
...         (?0)       #recursive call
...     )++          #end of non-capturing group, 1 or more times
...     \)           #literal )
... ''', flags=regex.X)

>>> lvlN.findall(eqn0)
```

```
['(b * c)', '(d / e)']

>>> lvln.findall(eqn1)
['((f+x)^y-42)', '((3-g)^z+2)']

>>> lvln.findall(eqn2)
['(b)', '((c))', '(((d)))']

>>> eqn3 = '(3+a) * ((r-2)*(t+2)/6) + 42 * (a(b(c(d(e)))))'
>>> lvln.findall(eqn3)
['(3+a)', '((r-2)*(t+2)/6)', '(a(b(c(d(e)))))']
```

Named character sets

A named character set is defined by a name enclosed between `[:` and `:` and has to be used within a character class `[]`, along with any other characters as needed. Using `[:^` instead of `[:` will negate the named character set. See [regular-expressions: POSIX Bracket](#) for full list, and refer to [pypi: regex](#) for notes on Unicode.

```
# similar to: r'\d+' or r'[0-9]+'
>>> regex.split(r'[:digit:]+', 'Sample123string42with777numbers')
['Sample', 'string', 'with', 'numbers']
# similar to: r'[a-zA-Z]+'
>>> regex.sub(r'[:alpha:]+', ':', 'Sample123string42with777numbers')
':123:42:777:'

# similar to: r'[\w\s]+'
>>> regex.findall(r'[:word][:space:]+', 'tea sea-pit sit-lean\tbean')
['tea sea', 'pit sit', 'lean\tbean']
# similar to: r'\S+'
>>> regex.findall(r'[:^space:]+', 'tea sea-pit sit-lean\tbean')
['tea', 'sea-pit', 'sit-lean', 'bean']

# words not surrounded by punctuation characters
>>> regex.findall(r'(?<[[:punct:]])\b\w+\b(?![[:punct:]])', 'tie. ink eat;')
['ink']
```

Set operations

Set operators can be used inside character class between sets. Mostly used to get intersection or difference between two sets, where one/both of them is a character range or a predefined character set. To aid in such definitions, you can use `[]` in nested fashion. The four operators, in increasing order of precedence, are:

- `||` union
- `~~` symmetric difference
- `&&` intersection
- `--` difference

```
# [^aeiou] will match any non-vowel character
# which means space is also a valid character to be matched
>>> regex.findall(r'\b[^aeiou]+\b', 'tryst glyph pity why')
['tryst glyph ', ' why']
# intersection or difference can be used here
# to get a positive definition of characters to match
>>> regex.findall(r'\b[a-z&[^aeiou]]+\b', 'tryst glyph pity why')
['tryst', 'glyph', 'why']

# [[a-l]~[g-z]] is same as [a-fm-z]
>>> regex.findall(r'\b[[a-l]~[g-z]]+\b', 'gets eat top sigh')
['eat', 'top']

# remove all punctuation characters except . ! and ?
>>> para = '"Hi", there! How *are* you? All fine here.'
>>> regex.sub(r'[:punct:]-[.!?]+', '', para)
'Hi there! How are you? All fine here.'
```



These set operators may get added to `re` module in future.

Unicode character sets

Similar to named character classes and escape sequence character sets, the `regex` module also supports `\p{}` construct that offers various predefined sets to work with Unicode strings. See [regular-expressions: Unicode](#) for details.

```
# extract all consecutive letters
>>> regex.findall(r'\p{L}+', 'fox:αλεπού,eagle:αετός')
['fox', 'αλεπού', 'eagle', 'αετός']
# extract all consecutive Greek letters
>>> regex.findall(r'\p{Greek}+', 'fox:αλεπού,eagle:αετός')
['αλεπού', 'αετός']

# extract all words
>>> regex.findall(r'\p{Word}+', 'φ0012,βτ_4,foo')
['φ0012', 'βτ_4', 'foo']

# delete all characters other than letters
# \p{^L} can also be used instead of \P{L}
>>> regex.sub(r'\P{L}+', '', 'φ0012,βτ_4,foo')
'φ00βτfoo'
```

Skipping matches

Sometimes, you want to change or extract all matches except particular matches. Usually, there are common characteristics between the two types of matches that makes it hard or

impossible to define RE only for the required matches. For example, changing field values unless it is a particular name, or perhaps don't touch double quoted values and so on. To use the skipping feature, define the matches to be ignored suffixed by `(*SKIP)(*FAIL)` and then define the matches required as part of alternation. `(*F)` can also be used instead of `(*FAIL)`.

```
# change lowercase whole words other than imp or rat
>>> words = 'tiger imp goat eagle rat'
>>> regex.sub(r'\b(?:imp|rat)\b(*SKIP)(*F)|[a-z]++', r'(\g<0>)', words)
'(tiger) imp (goat) (eagle) rat'

# change all commas other than those inside double quotes
>>> row = '1,"cat,12",nice,two,"dog,5"'
>>> regex.sub(r'"[^"]++"(*SKIP)(*F)|,', '|', row)
'1|"cat,12"|nice|two|"dog,5"'
```

`\m` and `\M` word anchors

`\m` and `\M` anchors match only the start and end of word respectively.

```
>>> regex.sub(r'\b', ': ', 'hi log_42 12b')
':hi: :log_42: :12b:'
>>> regex.sub(r'\m', ': ', 'hi log_42 12b')
':hi :log_42 :12b'
>>> regex.sub(r'\M', ': ', 'hi log_42 12b')
'hi: log_42: 12b:'

>>> regex.sub(r'\b..\b', r'[\g<0>]', 'I have 12, he has 2!')
'[I ]have [12][, ][he] has[ 2]!'
>>> regex.sub(r'\m..\M', r'[\g<0>]', 'I have 12, he has 2!')
'I have [12], [he] has 2!'
```

Overlapped matches

`findall` and `finditer` support `overlapped` optional argument. Setting it to `True` gives you overlapped matches.

```
>>> words = 'on vast ever road lane at peak'
>>> regex.findall(r'\b\w+ \w+\b', words)
['on vast', 'ever road', 'lane at']
>>> regex.findall(r'\b\w+ \w+\b', words, overlapped=True)
['on vast', 'vast ever', 'ever road', 'road lane', 'lane at', 'at peak']

>>> regex.findall(r'\w{2}', 'apple', overlapped=True)
['ap', 'pp', 'pl', 'le']
```

regex.REVERSE flag

The `regex.R` or `regex.REVERSE` flag will result in right-to-left processing instead of the usual left-to-right order.

```
>>> words = 'par spare lion part cool'

# replaces first match
>>> regex.sub(r'par', 'co', words, count=1)
'co spare lion part cool'

# replaces last match
>>> regex.sub(r'par', 'co', words, count=1, flags=regex.R)
'par spare lion cot cool'

>>> regex.findall(r'(?r)\w+', words)
['cool', 'part', 'lion', 'spare', 'par']
```

\X vs dot metacharacter

Some characters have more than one codepoint. These are handled in Unicode with **grapheme clusters**. The dot metacharacter will only match one codepoint at a time. You can use `\X` to match any character (including newline), even if it has multiple codepoints.

```
>>> [c.encode('unicode_escape') for c in 'ğ']
[b'g', b'\\u0308']

>>> regex.sub(r'a.e', 'o', 'cağed')
'cağed'
>>> regex.sub(r'a..e', 'o', 'cağed')
'cod'
>>> regex.sub(r'a\Xe', 'o', 'cağed')
'cod'

# \X will match newline character as well
>>> regex.sub(r'e.a', 'ea', 'he\nat', flags=regex.S)
'heat'
>>> regex.sub(r'e\Xa', 'ea', 'he\nat')
'heat'
```

Cheatsheet and Summary

Note	Description
<code>pypi: regex</code>	third party module, has lots advanced features
	default is <code>VERSION0</code> which is compatible with <code>re</code> module
<code>(?V1)</code>	inline flag to enable version 1 for <code>regex</code> module
	<code>regex.DEFAULT_VERSION=regex.VERSION1</code> can also be used
	<code>(?V0)</code> or <code>regex.VERSION0</code> to get back default version

Note	Description
possessive	enabled by appending <code>+</code> to greedy quantifier like greedy, but no backtracking
<code>(?>pat)</code>	atomic grouping, similar to possessive quantifier
<code>(?N)</code>	subexpression call for Nth capture group
<code>(?&name)</code>	subexpression call for named capture group
	subexpression call is similar to functions, recursion also possible <code>r'\((?:[^()]+\ (?0))+\)'</code> matches nested sets of parentheses
<code>pat\K</code>	<code>pat</code> won't be part of matching portion <code>\K</code> is used similar to positive lookbehind <code>regex</code> module allows variable length lookbehinds
<code>\G</code>	restricts matching from start of string like <code>\A</code> continues matching from end of match as new anchor until it fails <code>regex.findall(r'\G\d+-?', '12-34 42')</code> gives <code>['12-', '34']</code>
<code>[[:digit:]]</code>	named character set for <code>\d</code>
<code>[[:^digit:]]</code>	to indicate <code>\D</code>
	See regular-expressions: POSIX Bracket for full list
set operations	feature for character classes, nested <code>[]</code> allowed <code> </code> union, <code>~</code> symmetric difference <code>&</code> intersection, <code>--</code> difference <code>[[:punct:]]--[.!?]</code> punctuation except <code>.</code> , <code>!</code> and <code>?</code>
<code>\p{}</code>	Unicode character sets provided by <code>regex</code> module see regular-expressions: Unicode for details
<code>\P{L}</code> or <code>\p{^L}</code>	match characters other than <code>\p{L}</code> set
<code>pat(*SKIP)(*F)</code>	ignore text matched by <code>pat</code> <code>"[^"]++>(*SKIP)(*F) ,"</code> will match <code>,</code> but not inside double quoted pairs
<code>\m</code> and <code>\M</code>	anchors for start and end of word respectively
overlapped	set as <code>True</code> to match overlapping portions
<code>regex.R</code>	<code>REVERSE</code> flag to match from right-to-left
<code>\X</code>	matches any character even if it has multiple codepoints <code>\X</code> will also match newline characters by default whereas <code>.</code> requires <code>re.S</code> flag to match newline character

There's lots and lots of features provided by `regex` module. Some of them have not been covered in this chapter — for example, fuzzy matching and `splititer`. See [pypi: regex](#) for details and examples. For those familiar with Perl style regular expressions, this module offers easier transition compared to `re` module.

Exercises

a) Filter all elements whose first non-whitespace character is not a `#` character. Any element made up of only whitespace characters should be ignored as well.

```
>>> items = ['    #comment', '\t\napple #42', '#oops', 'sure', 'no#1', '\t\r\f']
```

```
##### add your solution here
['\t\napple #42', 'sure', 'no#1']
```

b) Replace sequences made up of words separated by `:` or `.` by the first word of the sequence and the separator. Such sequences will end when `:` or `.` is not followed by a word character.

```
>>> ip = 'wow:Good:2_two:five: hi bye kite.777.water.'

##### add your solution here
'wow: hi bye kite.'
```

c) The given list of strings has fields separated by `:` character. Delete `:` and the last field if there is a digit character anywhere before the last field.

```
>>> items = ['42:cat', 'twelve:a2b', 'we:be:he:0:a:b:bother']

##### add your solution here
['42', 'twelve:a2b', 'we:be:he:0:a:b']
```

d) Extract all whole words unless they are preceded by `:` or `<=>` or `----` or `#`.

```
>>> ip = '::very--at<=>row|in.a_b#b2c=>lion----east'

##### add your solution here
['at', 'in', 'a_b', 'lion']
```

e) The given input string has fields separated by `:` character. Extract all fields if the previous field contains a digit character.

```
>>> ip = 'vast:a2b2:ride:in:awe:b2b:3list:end'

##### add your solution here
['ride', '3list', 'end']
```

f) The given input string has fields separated by `:` character. Delete all fields, including the separator, unless the field contains a digit character. Stop deleting once a field with digit character is found.

```
>>> row1 = 'vast:a2b2:ride:in:awe:b2b:3list:end'
>>> row2 = 'um:no:low:3e:s4w:seer'

>>> pat = regex.compile() ##### add your solution here

>>> pat.sub('', row1)
'a2b2:ride:in:awe:b2b:3list:end'
>>> pat.sub('', row2)
'3e:s4w:seer'
```

g) For the given input strings, extract `if` followed by any number of nested parentheses. Assume that there will be only one such pattern per input string.

```
>>> ip1 = 'for (((i*3)+2)/6) if(3-(k*3+4)/12-(r+2/3)) while()'
>>> ip2 = 'if+while if(a(b)c(d(e(f)1)2)3) for(i=1)'
```

```
>>> pat = regex.compile() ##### add your solution here

>>> pat.search(ip1)[0]
'if(3-(k*3+4)/12-(r+2/3))'

>>> pat.search(ip2)[0]
'if(a(b)c(d(e(f)1)2)3)'
```

h) Read about `POSIX` flag from <https://pypi.org/project/regex/>. Is the following code snippet showing the correct output?

```
>>> words = 'plink incoming tint winter in caution sentient'

>>> change = regex.compile(r'int|in|ion|ing|inco|inter|ink', flags=regex.POSIX)

>>> change.sub('X', words)
'plX XmX tX wX X cautX sentient'
```

i) Extract all whole words for the given input strings. However, based on user input `ignore`, do not match words if they contain any character present in the `ignore` variable.

```
>>> s1 = 'match after the last newline character'
>>> s2 = 'and then you want to test'

>>> ignore = 'aty'
>>> regex.findall() ##### add your solution here for s1
['newline']
>>> regex.findall() ##### add your solution here for s2
[]

>>> ignore = 'esw'
>>> regex.findall() ##### add your solution here for s1
['match']
>>> regex.findall() ##### add your solution here for s2
['and', 'you', 'to']
```

j) Retain only punctuation characters for the given strings (generated from codepoints). Use Unicode character set definition for punctuation for solving this exercise.

```
>>> s1 = ''.join(chr(c) for c in range(0, 0x80))
>>> s2 = ''.join(chr(c) for c in range(0x80, 0x100))
>>> s3 = ''.join(chr(c) for c in range(0x2600, 0x27ec))

>>> pat = regex.compile() ##### add your solution here

>>> pat.sub('', s1)
'!"#$%&\'()*,-./:;?@[\\]_{}`'

>>> pat.sub('', s2)
'¡¢£«¬®¯°±²³´µ¶·¸¹º»¼½¾¿'

>>> pat.sub('', s3)
'(){}<>[]^_`{|}~¡¢£¥¦§¨ª«¬®¯°±²³´µ¶·¸¹º»¼½¾¿'
```

k) For the given **markdown** file, replace all occurrences of the string `python` (irrespective of case) with the string `Python`. However, any match within code blocks that start with whole line ````python` and end with whole line ````` shouldn't be replaced. Consider the input file to be small enough to fit memory requirements.

Refer to [github: exercises folder](#) for files `sample.md` and `expected.md` required to solve this exercise.

```
>>> ip_str = open('sample.md', 'r').read()
>>> pat = regex.compile()          ##### add your solution here
>>> with open('sample_mod.md', 'w') as op_file:
...     ##### add your solution here
...
305
>>> assert open('sample_mod.md').read() == open('expected.md').read()
```

l) For the given input strings, construct a word that is made up of last characters of all the words in the input. Use last character of last word as first character, last character of last but one word as second character and so on.

```
>>> s1 = 'knack tic pi roar what'
>>> s2 = '42;rod;t2t2;car'

>>> pat = regex.compile()          ##### add your solution here

##### add your solution here for s1
'trick'
##### add your solution here for s2
'r2d2'
```

m) Replicate `str.rpartition` functionality with regular expressions. Split into three parts based on last match of sequences of digits, which is `777` and `12` for the given input strings.

```
>>> s1 = 'Sample123string42with777numbers'
>>> s2 = '12apples'

##### add your solution here for s1
['Sample123string42with', '777', 'numbers']
##### add your solution here for s2
['', '12', 'apples']
```

n) Read about fuzzy matching on <https://pypi.org/project/regex/>. For the given input strings, return `True` if they are exactly same as `cat` or there is exactly one character difference. Ignore case when comparing differences. For example, `Ca2` should give `True`. `act` will be `False` even though the characters are same because position should be maintained.

```
>>> pat = regex.compile()          ##### add your solution here

>>> bool(pat.fullmatch('CaT'))
True
>>> bool(pat.fullmatch('scat'))
False
>>> bool(pat.fullmatch('ca.'))
```

```
True
>>> bool(pat.fullmatch('ca#'))
True
>>> bool(pat.fullmatch('c#t'))
True
>>> bool(pat.fullmatch('at'))
False
>>> bool(pat.fullmatch('act'))
False
>>> bool(pat.fullmatch('2a1'))
False
```

Gotchas

RE can get quite complicated and cryptic a lot of the times. But sometimes, if something is not working as expected, it could be because of quirky corner cases.

Escape sequences

Some RE engines match character literally if an escape sequence is not defined. Python raises an exception for such cases. Apart from sequences defined for RE (for example `\d`), these are allowed: `\a \b \f \n \N \r \t \u \U \v \x \\` where `\b` means backspace only in character classes and `\u \U` are valid only in Unicode patterns.

```
>>> bool(re.search(r'\t', 'cat\tdog'))
True

>>> bool(re.search(r'\c', 'cat\tdog'))
re.error: bad escape \c at position 0
```

Only octal escapes are allowed inside raw strings in replacement section. If you are otherwise not using `\` character, then using normal strings in replacement section is preferred as it will also allow `\x` and unicode escapes.

```
>>> re.sub(r',', r'\x7c', '1,2')
re.error: bad escape \x at position 0

>>> re.sub(r',', r'\174', '1,2')
'1|2'

>>> re.sub(r',', '\x7c', '1,2')
'1|2'
```

Line anchors with `\n` as last character

There is an additional start/end of line match after last newline character if line anchors are used as standalone pattern. End of line match after newline is straightforward to understand as `$` matches both end of line and end of string.

```
# note the use of inline flag to enable . to match newlines
>>> print(re.sub(r'(?m)^\s', 'foo ', '1\n2\n'))
foo 1
foo 2
foo

>>> print(re.sub(r'(?m)$\s', ' baz', '1\n2\n'))
1 baz
2 baz
baz
```

Zero length matches

How much does `*` or `*+` match? See also [regular-expressions: Zero-Length Matches](#).

```
# there is an extra empty string match at end of matches
>>> re.sub(r'^,*', r'\g<0>', ',cat,tiger')
'{}, {cat} {}, {tiger} {}'
>>> regex.sub(r'^,*+', r'\g<0>', ',cat,tiger')
'{}, {cat} {}, {tiger} {}'

# use lookarounds as a workaround
>>> re.sub(r'(?![^,])^,*', r'\g<0>', ',cat,tiger')
'{}, {cat}, {tiger}'
```

Capture group with quantifiers

Referring to text matched by a capture group with a quantifier will give only the last match, not entire match. Use a non-capturing group inside a capture group to get the entire matched portion.

```
>>> re.sub(r'\A([^,]+){3}([^,]+)', r'\1(\2)', '1,2,3,4,5,6,7')
'3,(4),5,6,7'
>>> re.sub(r'\A(?:[^,]+){3}([^,]+)', r'\1(\2)', '1,2,3,4,5,6,7')
'1,2,3,(4),5,6,7'

# as mentioned earlier, findall can be useful for debugging purposes
>>> re.findall(r'([^\,]+){3}', '1,2,3,4,5,6,7')
['3,', '6,']
>>> re.findall(r'(?:[^\,]+){3}', '1,2,3,4,5,6,7')
['1,2,3,', '4,5,6,']
```

Converting re to regex module

When using `flags` options with `regex` module, the constants should also be used from `regex` module. A typical workflow shown below:

```
# Using re module, unsure if a feature is available
>>> re.findall(r'[[:word:]]+', 'fox:αλεπού,eagle:αετός', flags=re.A)
<stdin>:1: FutureWarning: Possible nested set at position 1
[]

# Convert re to regex: oops, output is still wrong
>>> regex.findall(r'[[:word:]]+', 'fox:αλεπού,eagle:αετός', flags=re.A)
['fox', 'αλεπού', 'eagle', 'αετός']

# Finally correct solution, the constant had to be changed as well
>>> regex.findall(r'[[:word:]]+', 'fox:αλεπού,eagle:αετός', flags=regex.A)
['fox', 'eagle']
```

```
# or, use inline flags to avoid these shenanigans
>>> regex.findall(r'(?a)[[:word:]]+', 'fox:αλεπού,eagle:αετός')
['fox', 'eagle']
```

Optional arguments syntax

Speaking of `flags`, try to always use it as keyword argument. Using it as positional argument leads to a common mistake between `re.findall` and `re.sub` due to difference in placement. Their syntax, as per the docs, is shown below:

```
re.findall(pattern, string, flags=0)

re.sub(pattern, repl, string, count=0, flags=0)
```

Here's an example:

```
>>> +re.I
2

# works because flags is the only optional argument for findall
>>> re.findall(r'key', 'KEY portkey oKey Keyed', re.I)
['KEY', 'key', 'Key', 'Key']

# no error, because re.I has a value of 2
# this is same as count=2
>>> re.sub(r'key', 'X', 'KEY portkey oKey Keyed', re.I)
'KEY portX oKey Keyed'

# correct use of keyword argument
>>> re.sub(r'key', 'X', 'KEY portkey oKey Keyed', flags=re.I)
'X portX oX Xed'
```

Summary

Hope you have found Python regular expressions an interesting topic to learn. Sooner or later, you'll need to use them if you are doing plenty of text processing tasks. At the same time, knowing when to use normal string methods and knowing when to reach for other text parsing modules like `json` is important. Happy coding!

Further Reading

Note that some of these resources are not specific to Python, so use them with caution and check if they apply to Python's syntax and features.

- [docs.python: Regular Expression HOWTO](#)
- [stackoverflow: python regex](#)
- [CommonRegex](#) — collection of common regular expressions
- [Awesome Regex](#) — curated collection of libraries, tools, frameworks and software
- [Generate strings that match a given regular expression](#)
- [stackoverflow: regex FAQ](#)
 - [stackoverflow: regex tag](#) is a good source of exercise questions
- [rexegg](#) — tutorials, tricks and more
- [regular-expressions](#) — tutorials and tools
- [regexcrossword](#) — tutorials and puzzles, only supports JavaScript flavor
- [regex101](#) — visual aid and online testing tool for regular expressions, select flavor as Python before use
- [debuggex](#) — railroad diagrams for regular expressions, select flavor as Python before use
- [swtch](#) — stuff about regular expression implementation engines

Here's some links for specific topics:

- [rexegg: best regex trick](#)
- [regular-expressions: matching numeric ranges](#)
- [regular-expressions: Continuing at The End of The Previous Match](#)
- [regular-expressions: Zero-Length Matches](#)
- [stackoverflow: Greedy vs Reluctant vs Possessive Quantifiers](#)
- [cloudflare: Details of the Cloudflare outage on July 2, 2019](#) — see appendix for details about CPU exhaustion caused due to regular expression backtracking