localhost

Example (or, parentheses):

- What does `EECS 280|398` match?
- What does `EECS (280|398)` match?

?

11.1

regex101.com

regex101

REGULAR EXPRESSION

2 matches (23 steps, 60µs)

/ EECS (280|398) / gm

TEST STRING

last semester i took EECS 280 and now I'm in EECS 398

colored green because capturing group (relevant for re.findall())

1:48 – match 2

localhost

Example (closure, parentheses):

- What does `eecs*` match?
- What does `(eecs)*` match?

?

12.1

regex101.com

regex101

REGULAR EXPRESSION

2 matches (12 steps, 55µs)

/ eecs* / gm

0 or more of the previous character

TEST STRING

i•am•in•eecsssssssssssssssss•1234↵
and•my•friend•is•in•eec•15

2:27

localhost

Example (character classes, at least one):
`[A–E]+` is just shortform for `(A|B|C|D|E)(A|B|C|D|E)*`.

17.1

regex101.com

regex101

REGULAR EXPRESSION

6 matches (12 steps, 55µs)

/ [A-E2-7x-z+]+ / gm

TEST STRING

ABRAC234532453ADARBRA91445x2y2X+

special characters are usually escaped by default in [square brackets]

1:33

Example (number of occurrences): What does `tri{3, 5}` match? Does it match `'triiiiii'`?

?

20 . 1

regex101.com

regex101

REGULAR EXPRESSION

3 matches (8 steps, 35µs)

/ [0-9]{3,5} / gm

TEST STRING

ABRAC234532453ADARBRA91445x2y2X+-

matches 3-5 consecutive digits

regex is GREEDY: finds the longest possible matches.

1:34

localhost

escape character	`umich\.edu`	`'umich.edu'`	`'umich!edu'`
beginning of line	`^ark`	`'ark two'` `'ark o ark'`	`'dark'`
end of line	`ark$`	`'dark'` `'ark o ark'`	`'ark two'`
zero or one	`cat?`	`'ca'` `'cat'`	`'cart'` (matches `'ca'` only)
built-in character classes*	`\w+` `\d+`	`'billy'` `'231231'`	`'this person'` `'858 people'`
character class negation	`[^a-z]+`	`'WOLVERINE551'` `'1721$$'`	`'porch'` `'billy.edu'`

*_**Note**_: in Python's implementation of regex,

- `\d` refers to digits.
- `\w` refers to alphanumeric characters (`[A-Z][a-z][0-9]_`). **Whenever we say "alphanumeric" in an assignment, we're referring to `\w`!**
- `\s` refers to whitespace.
- `\b` is a word boundary.

?

24

regex101.com

regex101

REGULAR EXPRESSION

/ insert your regular expression here / gm

TEST STRING

insert your test string her

localhost

Example (escaping):

- What does `he.` match?
- What does `he\.` match?
- What does `(734)` match?
- What does `\(734\)` match?

?

25 . 1

regex101.com

regex101

REGULAR EXPRESSION

1 match (10 steps, 65µs)

/ \(734\) \d{3} \d{4} / gm

TEST STRING

hey there i go to michigan↵
my phone number is↵
(734) 764 1000

escaped, so treated as literals.

3:15

Raw strings

regex \b : word boundary
Python \b : backspace

When using regular expressions in Python, it's a good idea to use **raw strings**, denoted by an `r` before the quotes, e.g. `r'exp'`.

In [7]:
```
1 re.findall('\bcat\b', 'my cat is hungry')
```

Out[7]: []

In [8]:
```
1 re.findall(r'\bcat\b', 'my cat is hungry')
```

raw string

Out[8]: ['cat']

In [9]:
```
1 # Huh?
2 print('\bcat\b')
```

```
ca
```

15]:
```
# A regex that matches strings with two of the same vowel followed by
# We only want to capture the digits, but...
re.findall(r'(aa|ee|ii|oo|uu)(\d{3})', 'eeoo124')
```

15]:
```
[('oo', '124')]
```

[1 tuple with 2 elts]

To specify that we **don't** want to capture a particular group, u
at the start.

`?:` specifies a **non-capturing group**.

"don't capture this stuff, just match"

16]:
```
re.findall(r'(?:aa|ee|ii|oo|uu)(\d{3})', 'eeoo124')
```

16]:
```
['124']
```

[1 string]

17]:
```
# A regex that matches strings with two of the same vowel followed by
# We only want to capture the digits, but...
re.findall(r'(aa|ee|ii|oo|uu)(\d{3})', '''
check out these codes

eeoo124
xxaa519
hey there ee500
''')
```

17]:
```
[('oo', '124'), ('aa', '519'), ('ee', '500')]
```

To specify that we **don't** want to capture a particular group, u

at the start.

`?:` specifies a **non-capturing group**.

Out[24]: []

- We can use the Series `str.findall` method, with the regular expression above, to extract hashtags out of each tweet in `tweets['text']`.

In [28]:
```
1 tweets['text'].head()
```

Out[28]:
```
0    The Best Exercise To Lose Belly Fat In 2 weeks...
1    RT @Philanthropy: Dozens of 'hate groups' have...
2    Artificial intelligence can find, map poverty,...
3    Uber balks at rules proposed by world's busies...
4    RT @dirtroaddiva1: #IHatePokemonGoBecause he  ...
Name: text, dtype: object
```

In [26]:
```
1 tags = tweets['text'].str.findall(r'#(\w+)')
2 tags.head()
```

Out[26]:
```
0    [Exercise, LoseBellyFat, CatTV, TeenWolf]
1                                           []
2                                       [tech]
3                                       [news]
4       [IHatePokemonGoBecause, PokesAreJokes]
Name: text, dtype: object
```