

# Python Basic Assignment 7

## 1. What is the name of the feature responsible for generating Regex objects?

`re.compile()` is the feature responsible for generation of Regex objects.

```
In [1]: import re
x = re.compile("random pattern")
type(x)
print(x)
```

```
re.compile('random pattern')
```

## 2. Why do raw strings often appear in Regex objects?

Regular expressions use the backslash character ('\') to indicate special forms (Metacharacters) or to allow special characters (special sequences) to be used without invoking their special meaning. This collides with Python's usage of the same character for the same purpose in string literals. Hence, Raw strings are used (e.g. `r"\n"`) so that backslashes do not have to be escaped.

## 3. What is the return value of the `search()` method?

The return value of `re.search(pattern,string)` method is a match object if the pattern is observed in the string else it returns a `None`

```
In [2]: import re
match = re.search('i', 'Nikhilesh Singh', flags=re.IGNORECASE)
print('Output:', match)
match = re.search('X', 'Nikhilesh Singh', flags=re.IGNORECASE)
print('Output:', match)
```

```
Output: <re.Match object; span=(1, 2), match='i'>
```

```
Output: None
```

## 4. From a Match item, how do you get the actual strings that match the pattern?

For Matched items `group()` methods returns actual strings that match the pattern

```
In [5]: import re
match = re.search('singh','Nikhilesh Singh', flags=re.IGNORECASE)
print('Output:',match.group())
```

Output: Singh

## 5. In the regex which created from the `r'(\d\d\d)-(\d\d\d-\d\d\d\d)'`, what does group zero cover? Group 2? Group 1?

In the Regex `r'(\d\d\d)-(\d\d\d-\d\d\d\d)'` the zero group covers the entire pattern match where as the first group cover `(\d\d\d)` and the second group cover `(\d\d\d-\d\d\d\d)`

```
In [6]: # Example Program
import re
phoneNumRegex = re.compile(r'(\d\d\d)-(\d\d\d-\d\d\d\d)')
mo = phoneNumRegex.search('My number is 415-555-4242.')
print(mo.groups()) # Prints all groups in a tuple format
print(mo.group()) # Always returns the fully matched string
print(mo.group(1)) # Returns the first group
print(mo.group(2)) # Returns the second group
```

```
('415', '555-4242')
415-555-4242
415
555-4242
```

## 6. In standard expression syntax, parentheses and intervals have distinct meanings. How can you tell a regex that you want it to fit real parentheses and periods?

The `\.` `\(` and `\)` escape characters in the raw string passed to `re.compile()` will match actual parenthesis characters.

```
In [7]: # Example Program
import re
phoneNumRegex = re.compile(r'\((\d\d\d)\) (\d\d\d-\d\d\d\d)')
mo = phoneNumRegex.search('My phone number is (415) 555-4242.')
print(mo.group())
```

```
(415) 555-4242
```

## 7. The `findall()` method returns a string list or a list of string tuples. What causes it to return one of the two options?

If the regex pattern has no groups, a list of strings matched is returned. if the regex pattern has groups, a list of tuple of strings is returned.

```
In [8]: import re
phoneNumRegex = re.compile(r'(\d\d\d) (\d\d\d-\d\d\d\d)')
mo = phoneNumRegex.findall('My phone number is (415) 555-4242.')
print(mo)

# Example Program
import re
phoneNumRegex = re.compile(r'\d{3}-\d{3}-\d{4}')
mo = phoneNumRegex.findall('My number is 415-555-4242.')
print(mo) # Prints all groups in a tuple format

[('(415)', '555-4242')]
['415-555-4242']
```

## 8. In standard expressions, what does the | character mean?

In Standard Expressions | means OR operator.

## 9. In regular expressions, what does the ? character stand for?

In regular Expressions, ? characters represents zero or one match of the preceeding group.

```
In [10]: import re
match_1 = re.search("Bat(wo)?man", "Batman returns")
print(match_1)
match_2 = re.search("Bat(wo)?man", "Batwoman returns")
print(match_2)

<re.Match object; span=(0, 6), match='Batman'>
<re.Match object; span=(0, 8), match='Batwoman'>
```

## 10. In regular expressions, what is the difference between the + and \* characters?

In Regular Expressions, \* Represents Zero ore more occurances of the preceeding group, whereas + represents one or more occurances of the preceeding group.

```
In [11]: import re
match_1 = re.search("Bat(wo)*man", "Batman returns")
print(match_1)
match_2 = re.search("Bat(wo)+man", "Batman returns")
print(match_2)

<re.Match object; span=(0, 6), match='Batman'>
None
```

## 11. What is the difference between {4} and {4,5} in regular expression?

{4} means that its preceeding group should repeat 4 times. where as {4,5} means that its preceeding group should repeat minimum 4 times and maximum 5 times inclusively

```
In [12]: import re
haRegex = re.compile(r'(Ha){3}')
mo1 = haRegex.search('HaHaHa')
mo2 = haRegex.search('Ha')
print(mo1.group())
print(mo2)
```

HaHaHa  
None

## 12. What do you mean by the \d, \w, and \s shorthand character classes signify in regular expressions?

\d, \w and \s are special sequences in regular expressions in python:

1. \w – Matches a word character equivalent to [a-zA-Z0-9\_]
2. \d – Matches digit character equivalent to [0-9]
3. \s – Matches whitespace character (space, tab, newline, etc.)

## 13. What do means by \D, \W, and \S shorthand character classes signify in regular expressions?

\D, \W and \S are special sequences in regular expressions in python:

1. \W – Matches any non-alphanumeric character equivalent to [^a-zA-Z0-9\_]
2. \D – Matches any non-digit character, this is equivalent to the set class [^0-9]
3. \S – Matches any non-whitespace character

## 14. What is the difference between .\* and .\*?

.\* is a Greedy mode, which returns the longest string that meets the condition. Whereas .\*? is a non greedy mode which returns the shortest string that meets the condition.

## 15. What is the syntax for matching both numbers and lowercase letters with a character class?

The Syntax is Either [a-z0-9] or [0-9a-z]

## 16. What is the procedure for making a normal expression in regex case insensitive?

We can pass `re.IGNORECASE` as a flag to make a normal expression case insensitive

## 17. What does the `.` character normally match? What does it match if `re.DOTALL` is passed as 2nd argument in `re.compile()`?

Dot `.` character matches everything in input except newline character `\n`. By passing `re.DOTALL` as a flag to `re.compile()`, you can make the dot character match all characters, including the newline character.

## 18. If `numReg = re.compile(r'\d+')`, what will `numRegex.sub('X', '11 drummers, 10 pipers, five rings, 4 hen')` return?

The Output will be 'X drummers, X pipers, five rings, X hen'

```
In [14]: import re
numReg = re.compile(r'\d+')
numReg.sub('X', '11 drummers, 10 pipers, five rings, 4 hen')
```

```
Out[14]: 'X drummers, X pipers, five rings, X hen'
```

## 19. What does passing `re.VERBOSE` as the 2nd argument to `re.compile()` allow to do?

`re.VERBOSE` will allow to add whitespace and comments to string passed to `re.compile()`.

```
In [15]: # Without Using VERBOSE
regex_email = re.compile(r'^([a-z0-9_\-\.]+)@([0-9a-z\-\_\.]+)\.([a-z\-\_]{2, 6})$',

# Using VERBOSE
regex_email = re.compile(r"""
    ^([a-z0-9_\-\.]+)           # local Part like us
    @                           # single @ sign
    ([0-9a-z\-\_\.]+)          # Domain name like g
    \.                          # single Dot .
    ([a-z]{2,6})$              # Top level Domain
""", re.VERBOSE | re.IGNORECASE)
```

## 20. How would you write a regex that match a number with comma for every three digits? It must match the given following:

'42','1,234', '6,368,745' but not the following: '12,34,567' (which has only two digits between the commas) '1234' (which lacks commas)

```
In [16]: import re
pattern = r'^\d{1,3}(\,\d{3})*$'
pagex = re.compile(pattern)
for ele in ['42', '1,234', '6,368,745', '12,34,567', '1234']:
    print('Output:', ele, '->', pagex.search(ele))
```

Output: 42 -> <re.Match object; span=(0, 2), match='42'>  
 Output: 1,234 -> <re.Match object; span=(0, 5), match='1,234'>  
 Output: 6,368,745 -> <re.Match object; span=(0, 9), match='6,368,745'>  
 Output: 12,34,567 -> None  
 Output: 1234 -> None

## 21. How would you write a regex that matches the full name of someone whose last name is Watanabe? You can assume that the first name that comes before it will always be one word that begins with a capital letter. The regex must match the following:

```
pattern = r'[A-Z]{1}[a-z]*\sWatanabe'
```

```
In [17]: import re
pattern = r'[A-Z]{1}[a-z]*\sWatanabe'
namex = re.compile(pattern)
for name in ['Haruto Watanabe', 'Alice Watanabe', 'RoboCop Watanabe', 'haruto Wat
    print('Output: ', name, '->', namex.search(name))
```

Output: Haruto Watanabe -> <re.Match object; span=(0, 15), match='Haruto Watanabe'>  
 Output: Alice Watanabe -> <re.Match object; span=(0, 14), match='Alice Watanabe'>  
 Output: RoboCop Watanabe -> <re.Match object; span=(4, 16), match='Cop Watanabe'>  
 Output: haruto Watanabe -> None  
 Output: Mr. Watanabe -> None  
 Output: Watanabe -> None  
 Output: Haruto watanabe -> None

**22. How would you write a regex that matches a sentence where the first word is either Alice, Bob, or Carol; the second word is either eats, pets, or throws; the third word is apples, cats, or baseballs; and the sentence ends with a period? This regex should be case-insensitive. It must match the following:**

```
pattern = r'(Alice|Bob|Carol)\s(eats|pets|throws)\s(apples|cats|baseballs).'
```

```
In [19]: import re
pattern = r'(Alice|Bob|Carol)\s(eats|pets|throws)\s(apples|cats|baseballs)\.'
casex = re.compile(pattern, re.IGNORECASE)
for ele in ['Alice eats apples.', 'Bob pets cats.', 'Carol throws baseballs.', 'A
', 'ALICE THROWS FOOTBALLS.', 'Carol eats 7 cats.']:
    print('Output: ', ele, '->', casex.search(ele))
```

```
Output: Alice eats apples. -> <re.Match object; span=(0, 18), match='Alice e
ats apples.'>
```

```
Output: Bob pets cats. -> <re.Match object; span=(0, 14), match='Bob pets ca
ts.'>
```

```
Output: Carol throws baseballs. -> <re.Match object; span=(0, 23), match='Ca
rol throws baseballs.'>
```

```
Output: Alice throws Apples. -> <re.Match object; span=(0, 20), match='Alice
throws Apples.'>
```

```
Output: BOB EATS CATS. -> <re.Match object; span=(0, 14), match='BOB EATS CA
TS.'>
```

```
Output: RoboCop eats apples. -> None
```

```
Output: ALICE THROWS FOOTBALLS. -> None
```

```
Output: Carol eats 7 cats. -> None
```