Assignment 7 Solutions

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

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

In [1]:

```
import re
x = re.compile("some_random_pattern")
type(x)
print(x)
```

```
re.compile('some_random_pattern')
```

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

ANS: Regular expressions use the backslash character `('\')` to indicate special forms `(Metacharacters)` or to allow special characters `(speical 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?

ANS: 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','All Over Best Ineuron Full Stack Data Science Program', flags=re
print('Output:',match)
match = re.search('Z','All Over Best Ineuron Full Stack Data Science Program', flags=re
print('Output:',match)
```

```
Output: <re.Match object; span=(14, 15), match='I'>
Output: None
```

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

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

In [4]:

```
import re
match = re.search('All Over Best','All Over Best Ineuron Full Stack Data Science Progra
print('Output:',match.group())
```

```
Output: All Over Best
```

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?

ANS: 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 [8]:

```
# Example Program
import re
phoneNumRegex = re.compile(r'(\d\d\d)-(\d\d\d-\d\d\d\d)')
mo = phoneNumRegex.search('My number is 060-888-6615.')
print(mo.groups())
print(mo.group())
print(mo.group(1))
print(mo.group(2))
```

```
('060', '888-6615')
060-888-6615
060
888-6615
```

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?

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

In [12]:

```
# 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 (060) 888-6615.')
print(mo.group())
```

```
(060) 888-6615
```

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

ANS: 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 [13]:

```
import re
phoneNumRegex = re.compile(r'(\(\d\d\d\)) (\d\d\d-\d\d\d\d)')
mo = phoneNumRegex.findall('My phone number is (060) 888-6615')
print(mo)

import re
phoneNumRegex = re.compile(r'\d{3}-\d{3}-\d{4}')
mo = phoneNumRegex.findall('My number is 060-888-6615.')
print(mo)
```

```
[('(060)', '888-6615')]
['060-888-6615']
```

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

ANS: In Standard Expressions | means OR operator.

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

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

In [14]:

```
import re
match_1 = re.search("Super(wo)?man","Superman returns")
print(match_1)
match_2 = re.search("Super(wo)?man","Superwoman returns")
print(match_2)
```

```
<re.Match object; span=(0, 8), match='Superman'>
<re.Match object; span=(0, 10), match='Superwoman'>
```

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

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

In [15]:

```
import re
match_1 = re.search("Super(wo)*man","Superman returns")
print(match_1)
match_2 = re.search("Super(wo)+man","Superman returns")
print(match_2)
```

```
<re.Match object; span=(0, 8), match='Superman'>
None
```

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

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

In [16]:

```
import re
haRegex = re.compile(r'(Abhi){4}')
mo1 = haRegex.search('AbhiAbhiAbhiAbhi')
mo2 = haRegex.search('Abhi')
print(mo1.group())
print(mo2)
```

```
AbhiAbhiAbhiAbhi
None
```

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

ANS: \d, \w and \s are special sequences in regular expresssions 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?

ANS: \D, \W and \S are special sequences in regular expresssions 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 `.*`?

ANS: **`.*`** 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?

ANS: The Synatax is Either **`[a-z0-9]`** or **`[0-9a-z]`**

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

ANS: 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()?

ANS: Dot **`.`** character matches everything in input except newline character **`.`** . 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?

ANS: The Ouput will be `'X drummers, X pipers, five rings, X hen'`

In [22]:

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

Out[22]:

```
'X drummers, X pipers, five rings, X hen'
```

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

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

In [23]:

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

# Using VERBOSE
regex_email = re.compile(r"""
                            ^([a-z0-9_\.-]+)             # local Part like username
                            @                            # single @ sign
                            ([0-9a-z\.-]+)               # Domain name like google
                            \.                           # single Dot .
                            ([a-z]{2,6})$                # Top level Domain  like com/
                         """,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 [24]:

```
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:

`'Haruto Watanabe'`
`'Alice Watanabe'`
`'RoboCop Watanabe'`

but not the following:

`'haruto Watanabe'` (where the first name is not capitalized)
`'Mr. Watanabe'` (where the preceding word has a nonletter character)
`'Watanabe'` (which has no first name)
`'Haruto watanabe'` (where Watanabe is not capitalized)

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

In [25]:

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

```
Output:  Haruto Watanabe -> <re.Match object; span=(0, 15), match='Haruto Wa
tanabe'>
Output:  Alice Watanabe -> <re.Match object; span=(0, 14), match='Alice Wata
nabe'>
Output:  RoboCop Watanabe -> <re.Match object; span=(4, 16), match='Cop Wata
nabe'>
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:

```
'Alice eats apples.'
'Bob pets cats.'
'Carol throws baseballs.'
'Alice throws Apples.'
'BOB EATS CATS.'
```

but not the following:

```
'RoboCop eats apples.'
'ALICE THROWS FOOTBALLS.'
'Carol eats 7 cats.'
```

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

In [26]:

```
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.','Alice thro
,'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
eats apples.'>
Output:  Bob pets cats. -> <re.Match object; span=(0, 14), match='Bob pets c
ats.'>
Output:  Carol throws baseballs. -> <re.Match object; span=(0, 23), match='C
arol throws baseballs.'>
Output:  Alice throws Apples. -> <re.Match object; span=(0, 20), match='Alic
e throws Apples.'>
Output:  BOB EATS CATS. -> <re.Match object; span=(0, 14), match='BOB EATS C
ATS.'>
Output:  RoboCop eats apples. -> None
Output:  ALICE THROWS FOOTBALLS. -> None
Output:  Carol eats 7 cats. -> None
```

In [ ]:

```

```