Regular Expressions

About the Module

- What are Regular Expressions
- Use of Regular Expressions
- Types of Regular Expressions
- Examples

What are Regular Expressions

- Combination of special characters used to search patterns in the text objects
- Wild card expressions for matching, searching, and parsing strings
- Using regular expressions we can perform rule based text searching

```
/^[a-z0-9_-]{6,18}$/
```

Use of Regular Expressions

- Segmentation of words from sentences
- Segmentation of sentence from paragraphs
- Text Cleaning
- Information retrieval from large texts.

Types of Regular Expressions

abc…	Letters	{m}	m Repetitions
123…	Digits	{m,n}	m to n Repetitions
\d	Any Digit	*	Zero or more repetitions
\D	Any Non-digit character	+	One or more repetitions
.	Any Character	?	Optional character
\.	Period	\s	Any Whitespace
[abc]	Only a, b, or c	\S	Any Non-whitespace character
[^abc]	Not a, b, nor c	^…$	Starts and ends
[a-z]	Characters a to z	(…)	Capture Group
[0-9]	Numbers 0 to 9	(a(bc))	Capture Sub-group
\w	Any Alphanumeric character	(.*)	Capture all
\W	Any Non-alphanumeric character	(abc|def)	Matches abc or def

Regular Expression Functions

- match : finds the first occurrence of pattern in the string
- search : locates the pattern in the string
- findall : find all occurrences of patterns in the string
- sub : search and replace
- split : split the text by the given regular expression pattern

Examples

```
import re

string = "Tiger is the national animal of India"
pattern = "Tiger"

result = re.match(pattern, string).group(0)
>> Tiger
```

```
string = "The national animal of India is Tiger"
pattern = "Tiger"

result = re.search(pattern, string).group(0)
>> Tiger
```

```
string = "the national animal of India is tiger and the national sport of India is hockey"
pattern = "national"

re.findall(pattern, string)
>> ["national", "national"]
```

Examples

```
date_pattern = r' \d{2}-\d{2}-\d{4}'
text = “John 34-3456 12-05-2007, XYZ 56-4532 11-11-2011, ABC 67-8945 12-01-2009"
re.findall(date_pattern, text)
>> [' 12-05-2007', ' 11-11-2011', ' 12-01-2009']

line = 'asdf fjdk;afed,fjek,asdf,foo'
re.split(r'[;,\s]', line)
>> ['asdf', 'fjdk', 'afed', 'fjek', 'asdf', 'foo']

string = “cricket is a popular sport of India”
pattern = “India”
replacement = “the world”
re.sub(pattern, replacement, string)
>> “cricket is a popular sport of the world
```