Module 6: Regular Expressions

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 (or if a given regular expression matches a particular string, which comes down to the same thing).

The re-module in Python gives full support for regular expressions of Pearl style. The re module raises the re.error exception whenever an error occurs while implementing or using a regular expression.

Regular expressions can be concatenated to form new regular expressions; if *A* and *B* are both regular expressions, then *AB* is also a regular expression. In general, if a string *p* matches *A* and another string *q* matches *B*, the string *pq* will match AB. This holds unless *A* or *B* contain low precedence operations; boundary conditions between *A* and *B*; or have numbered group references. Thus, complex expressions can easily be constructed from simpler primitive expressions.

Regular expressions can contain both special and ordinary characters. Most ordinary characters, like 'A', 'a', or '0', are the simplest regular expressions; they simply match themselves. You can concatenate ordinary characters, so last matches the string 'last'.

Repetition operators or quantifiers (*, +, ?, {m,n}, etc) cannot be directly nested. This avoids ambiguity with the non-greedy modifier suffix ?, and with other modifiers in other implementations. To apply a second repetition to an inner repetition, parentheses may be used. For example, the expression (?:a{6})* matches any multiple of six 'a' characters.

Metacharacters

Every character in a Python RegEx is either a metacharacter or a regular character. A metacharacter has a special meaning, whereas a regular character matches itself.

Some of the basic metacharacters used in RegEx include:

- “^”

The ‘ ^ ’ character checks if the string starts with a particular word or character.

The ‘$‘ character checks if the string ends with a particular word or character.

The ‘ | ‘ character is used to check either/or condition.

The “+” matches one or more occurrences of a character in a string.

Special sequences

A special sequence is a ‘ \ ‘ symbol, followed by one of the particular characters. Some special sequences include:

The \A checks if the string starts with a particular character.

The \s sequence returns a match when the string contains white space characters.

The \d sequence checks if there are any digits in the given string.

The \Z sequence checks if the string ends with a particular word.

The \w sequence returns a match at every word character.