

Artificial Intelligence Masterclass

Regular Expressions (RegEx)

H.M. Samadhi Chathuranga Rathnayake

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU),
P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Introduction to RegEx

- A string of characters that creates a search pattern is known as a regex, or regular expression.
- RegEx can be used to determine whether a string includes a given search pattern.

RegEx Functions

Function	Description
<u>findall</u>	Returns a list containing all matches
<u>search</u>	Returns a <u>Match object</u> if there is a match anywhere in the string
<u>split</u>	Returns a list where the string has been split at each match
<u>sub</u>	Replaces one or many matches with a string

Metacharacters

Character	Description	Example
<code>[]</code>	A set of characters	<code>"[a-m]"</code>
<code>\</code>	Signals a special sequence (can also be used to escape special characters)	<code>"\d"</code>
<code>.</code>	Any character (except newline character)	<code>"he..o"</code>
<code>^</code>	Starts with	<code>"^hello"</code>
<code>\$</code>	Ends with	<code>"planet\$"</code>
<code>*</code>	Zero or more occurrences	<code>"he.*o"</code>
<code>+</code>	One or more occurrences	<code>"he.+o"</code>
<code>?</code>	Zero or one occurrences	<code>"he.?o"</code>
<code>{}</code>	Exactly the specified number of occurrences	<code>"he.{2}o"</code>
<code> </code>	Either or	<code>"falls stays"</code>

Special Sequences

Character	Description	Example
<code>\A</code>	Returns a match if the specified characters are at the beginning of the string	<code>"\AThe"</code>
<code>\b</code>	Returns a match where the specified characters are at the beginning or at the end of a word (the "r" in the beginning is making sure that the string is being treated as a "raw string")	<code>r"\bain"</code> <code>r"ain\b"</code>
<code>\B</code>	Returns a match where the specified characters are present, but NOT at the beginning (or at the end) of a word (the "r" in the beginning is making sure that the string is being treated as a "raw string")	<code>r"\Bain"</code> <code>r"ain\B"</code>
<code>\d</code>	Returns a match where the string contains digits (numbers from 0-9)	<code>"\d"</code>
<code>\D</code>	Returns a match where the string DOES NOT contain digits	<code>"\D"</code>
<code>\s</code>	Returns a match where the string contains a white space character	<code>"\s"</code>
<code>\S</code>	Returns a match where the string DOES NOT contain a white space character	<code>"\S"</code>
<code>\w</code>	Returns a match where the string contains any word characters (characters from a to Z, digits from 0-9, and the underscore <code>_</code> character)	<code>"\w"</code>
<code>\W</code>	Returns a match where the string DOES NOT contain any word characters	<code>"\W"</code>
<code>\Z</code>	Returns a match if the specified characters are at the end of the string	<code>"Spain\Z"</code>

Sets

Set	Description
[arn]	Returns a match where one of the specified characters (a , r , or n) is present
[a-n]	Returns a match for any lower case character, alphabetically between a and n
[^arn]	Returns a match for any character EXCEPT a , r , and n
[0123]	Returns a match where any of the specified digits (0 , 1 , 2 , or 3) are present
[0-9]	Returns a match for any digit between 0 and 9
[0-5][0-9]	Returns a match for any two-digit numbers from 00 and 59
[a-zA-Z]	Returns a match for any character alphabetically between a and z , lower case OR upper case
[+]	In sets, + , * , . , , () , \$, {} has no special meaning, so [+] means: return a match for any + character in the string