# Regular expressions

In windows, you can perform a search for all/any PDFs by using the command \*.pdf. The star is a regular expression that performs a wildcard search for all/any PDFs.

A regular expression is a type of notation that contains strings of characters that can be matched to the contents of a set.

Another regular expression may look like: a+b

In this example, the + means there are one or more of the proceeding element i.e. ab, aab, aaab etc

## What is a regular expression?

Regular expressions are a tool that enables programmers and computers to work with text patterns. They are used for example:

* To match patterns in text files (for example when searching for a particular word in a word processing program)
* By compliers to recognise the correct form of a variable name or the syntax of a statement (variables can NOT start with a number)
* By programmers to validate user input (for example to check that a postcode or an email address is in the correct format)

## What symbols used in regular expression do I need to know?

* | A vertical bar separates alternatives (Boolean OR)
* ? A question mark indicates that there are zero or one of the preceding element
* \*An asterisk indicates that there are zero or more of the preceding element
* + A superscript plus sign indicates that there is one or more of the preceding element

## How do they work?

|  |  |  |
| --- | --- | --- |
| Regular expression | Meaning | Matching string |
| (Edward) | (Eddie) | (Ed) | Boolean OR | Edward, Eddie, Ed |
| (D | d) is (c | k) | Brackets show scope  | “Disc”, “disc”, Disk”, “disk” |
| Dialog(ue)? | ? A question mark indicates that there are zero or one of the preceding element | Dialog, Dialogue |
| ab\* | \*An asterisk indicates that there are zero or more of the preceding element | a, ab, abb, abbb, ... |
| A+b | + A superscript plus sign indicates that there is one or more of the preceding element | ab, aab, aaab, ... |

## Question – complete this table

|  |  |
| --- | --- |
| Regular expression | Matching string |
| (Jamie) | (James) | (Jim) |  |
| (N | n) i (te |ght ) |  |
| Master(chef)? |  |
| bo\* |  |
| o+h | oh, ooh, oooh, ... |
|  | Davidson, David, Dave |
|  | ac, bc |
|  | Method, Methodman |
|  | P, po, poo, pooo, ... |
| a+h |  |
| col(o|ou)r  |  |
| ab\*a |  |
| 0+1+0 |  |

## Example 1

A regular language consists of all words beginning and ending in *a*, with zero or more instances of *b* in between e.g. aa, aba, abba, abbba.

Write a regular expression that describes this language and draw the corresponding finite state machine.

Answer: R = ab\*a. Note that the FSM is drawn with an outgoing transition from every state for every possible input symbol.



## Question

Write a regular expression to find all the occurrences of “color” or “colour” in a document.

## Question

Write a regular expression that matches any non-empty string that starts with zero or more “a”s, followed by one or more “b”s.

## Question

Which of the following strings is matched by the regular expression Sc(o+)(b|d)\*y

Scooby scoby scddy scobby scoobdbdbdy

## Example 1

Consider the FSM shown below, which has four states



This allows an empty string of the form ab, aabb and all combinations of thee such as abab, aabb, aababb.

The corresponding regular expression is: (a(ab)\*b)\*

## Question

Write the regular expression which represents the finite state machine shown below



## Question

Write the regular expression which represents the finite state machine shown below



## Question

Regular expressions can be used to search for strings. For example, de (f | g) \*h+ matches any string that starts with *de* and is followed by zero or more instances of either *f* or *g* followed by one or more instances of *h*.

1. Write regular expressions that will match:
2. Any string that starts with a letter *a*, ends with a letter *c* and has one or more occurrences of the letter *b* in the middle of it, i.e. the expression should match the strings abc, abbc, abbbc, and so on
3. Any string that starts with either a 0 or a 1, followed by zero or more occurrences of the digit 1, i.e. the expression should match the strings 0, 1, 01, 11, 011 and so on

1. Regular expressions can be used to search for strings.
2. For each of the following regular expressions, describe the set of strings that they would find.
3. a+b
4. a?b
5. (ab) \*
6. Which of the following strings will be accepted by the FSM shown below?



11001

01000

101111

000110