LECTURE 10

PRACTICAL REGULAR EXPRESSIONS

SUBJECTS

Practical notations that are often used with regular expression

Few practice exercises

PRACTICAL REGULAR EXPRESSIONS TRICKS

uOttawa
L'Université canadienne
Canada's university

We will see practical regular expressions tricks that are supported by most regex libraries

Remember, regular expressions are not only used in the context of compilers

- We often use them to extract information from text

<u>Example:</u> imagine looking in a log file that has been accumulating entries for the past two months for a particular error pattern

- Without regular expressions, this would be a tedious job
- Sooner or later, when you work in the industry, you will encounter such issues ➔ regular expressions will come in handy

MATCHING DIGITS

To match a single digit, as we have seen before, we can use the following regular expression:

[0-9]

Nonetheless, since matching a digit is a common operation, we can use the following notation:

\d

Slash is an escape character used to distinguish it from the letter d

Similarly, to match a non-digit character, we can use the notation:

\D

ALPHANUMERIC CHARACTERS

uOttawa
L'Université canadienne
Canada's university

To match an alphanumeric character, we can use the notation:

[a-zA-Z0-9]

Or we can use the following shortcut

\w

Similarly, we can represent any non-alphanumeric character as follows:

\W

MATCHING ANY LETTER CHARACTER

You can match any letter character in the English alphabet:

[a-zA-Z]

However, there are many letter characters defined in Unicode beyond the 26 letters of the English alphabet

- You can match those using:

  \p{L}

WILDCARD

A wildcard is defined to match any single character (letter, digit, whitespace …)

It is represented by the . (dot) character

Therefore, in order to match a dot, you have to use the escape character: \.

EXCLUSION

We have seen that [abc] is equivalent to (a | b | c)

But sometimes we want to match everything except a set of characters

To achieve this, we can use the notation: [^abc]

- This matches any single character other than a, b or c

This notation can also be used with abbreviated character classes

- [^a-z] matches any character other than a small letter

REPETITIONS

How can we match a letter or a string that repeats several times in a row:

- E.g. ababab

So far, we have implemented repetitions through three mechanisms:

- Concatenation: simply concatenate the string or character with itself (does not work if you do not know the exact number of repetitions)
- Kleene star closure: to match letters or strings repeated 0 or more times
- Positive closure: to match letters or strings repeated 1 or more times

REPETITIONS

We can also specify a range of how many times a letter or string can be repeated

Example, if we want to match strings of repetition of the letter a between 1 and 3 times, we can use the notation: a {1,3}

- Therefore, a {1,3} matches the strings
  - a
  - aa
  - aaa

We can also specify an exact number of repetitions instead of a range

- (ab) {3} matches the string ababab

OPTIONAL CHARACTERS

The concept of the optional character is somewhat similar to that of the kleene star

- The star operator matches **0 or more** instances of the operand
- The optional operator, denoted as ? (question mark), matches **0 or 1** instance of the operand

Example: the pattern ab?c will match either the strings "abc" or "ac" because the b is considered optional.

- This is also equivalent to ***a(b|ε)c***

WHITE SPACE

Often, we want to easily detect white spaces

- Either to remove them or to detect the beginning or end of words

Most common forms of whitespace used with regular expressions:

- Space _, the tab \t, the new line \n and the carriage return \r

A whitespace special character \s will match *any* of the specific whitespaces above

Similarly, you can match any non-white space character using the notation \S

MATCHING THE BEGINNING OR END OF A LINE

uOttawa
L'Université canadienne
Canada's university

You can specify that a regular expression match only the beginning or end of the line

If a caret (^) is at the beginning of the entire regular expression, it matches the beginning of a line

If a dollar sign ($) is at the end of the entire regular expression, it matches the end of a line

If an entire regular expression is enclosed by a caret and dollar sign (^this is a sentence$), it matches an entire line

MATCH ANY WORD BOUNDARY

uOttawa
L'Université canadienne
Canada's university

Sometimes it is important to match word boundaries

- Allows us to differentiate between strings that appear at the beginning of a word from those that appear in the middle

The beginning of a word is matched using \b

- For example: \bart matches the string "art" but not "st**art**"

The end of a word is also matched by \b

- For example: art\b matches the string "st**art**" but not "arts"

Similarly, the less useful \B matches the middle of a word

- For example: \Bart matches the string "st**art**" but not "art"

FEW EXERCISES - 1

uOttawa
L'Université canadienne
Canada's university

Given the sentence:

- Error, computer will now shut down…

Provide a regular expression that will match any word in the sentence

Answer:

[A-Za-z]+

or

\p{L}+

FEW EXERCISES - 2

Given the sentence:

- Error, computer will now shut down…

Provide a regular expression that will match all sequences of non-alphanumeric characters

Answer:

\W+

FEW EXERCISES - 3

Given the log file:

- [Sunday Feb. 2 2018] Program starting up
- [Monday Feb. 3 2018] Entered initialization phase
- [Tuesday Feb. 4 2018] Error 5: cannot open XML file
- [Thursday Feb. 6 2018] Warning 5: response time is too slow
- [Friday Feb. 7 2018] Error 9: major error occurred, system will shut down

Match any error or warning message that ends with the term "shut down"

Answer:

(Error|Warning).*(shut down)

FEW EXERCISES - 4

Given the log file:

- [Sunday Feb. 2 2018] Program starting up
- [Monday Feb. 3 2018] Entered initialization phase
- [Tuesday Feb. 4 2018] Error 5: cannot open XML file
- [Thursday Feb. 6 2018] Warning 5: response time is too slow
- [Friday Feb. 7 2018] Error 9: major error occurred, system will shut down

Match any Error or Warning between 1 and 6th February 2017

Answer:

\[\w+ Feb\. [1-6] 2018\] (Error|Warning)

FEW EXERCISES - 5

Specify the regular expression for passwords that must start by a letter followed by alphanumeric characters

The Password must consists of at least 8 characters and not more than 15 characters

\b[a-zA-Z]\w{7,14}\b

THANK YOU!

QUESTIONS?