REGULAR EXPRESSIONS FOR TRANSLATORS AND INTERPRETERS

28 March 2022

BEFORE THE WORKSHOP

Please install the latest version of Notepad++ from https://notepad-plus-plus.org/downloads/v8.3.3/ We will be using it for some of the exercises. If you cannot install it, you can use Microsoft Word instead.

It would be useful if in preparation for the workshop you open the following tabs in your browser so you can access them quickly:

- https://pollev.com/corasan for the various interactive exercises during the CPD. Providing a name is not compulsory, but it would be nice to add it.
- https://regexr.com/ for testing different regular expressions

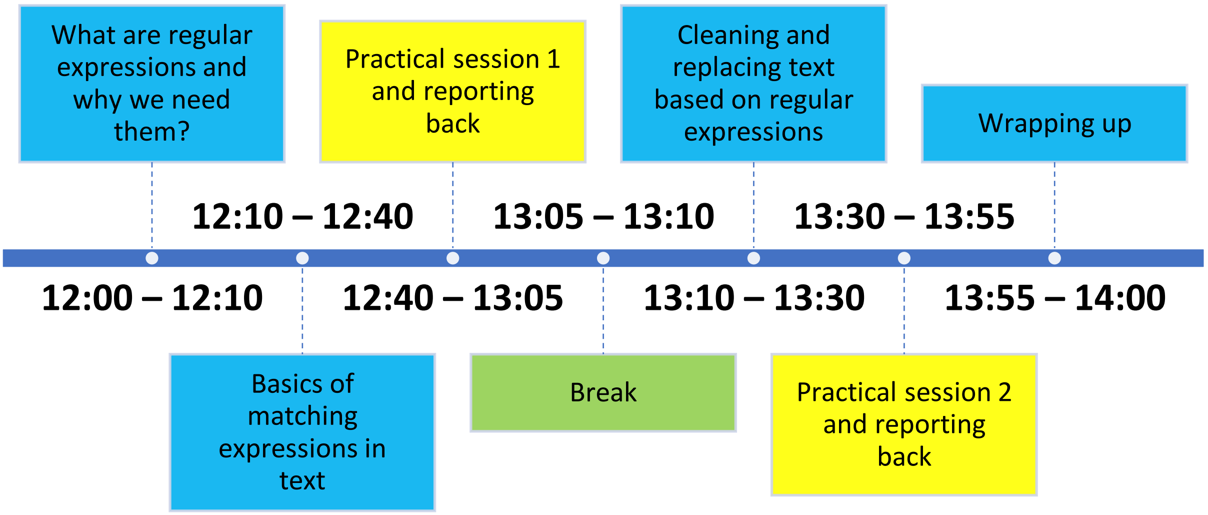
It would be useful if you could print this document or at least page 4 (the cheatsheet) as you will need it during the practical exercises. If you can't print it, make sure you have it open in a tab, so you can consult it quickly.

ZOOM LINK

In case you misplaced the zoom link, here it is: https://surrey-ac.zoom.us/j/94011039599?pwd=eWNQRFdCeW5OcGdNdDA5YjJNQkR1Zz09

Unless we experience some technical difficulties, we will start at 12pm (UK time) sharp. We have lots of stuff to cover.

PLANNED STRUCTURE FOR THE WORKSHOP



12:00 – 12:10 What are regular expressions and why we need them?

12:10 – 12:40 Basics of matching expressions in text

12:40 – 13:05 Practical session 1 and reporting back

13:05 – 13:10 Break

13:10 – 13:30 Cleaning and replacing text based on regular expressions

13:30 – 13:55 Practical session 2 and reporting back

13:55 – 14:00 Wrapping up

PRACTICAL SESSION 1

Write regular expressions which match the following expressions

- » Match both *color* and *colour.* How can you match both capitalised and lower-case words?
- » What kind of words the following expressions match?
    - `^[0-9]+\.[0-9]+$`
    - `[A-Z]+\$$`
    - `^[0-9]{4}$`
    - `^[0-9]+-[a-z]{3,5}$`
    - `^[a-z]{5,}-[a-z]{2,3}-[a-z]{,6}$`
    - `(ed|ing)$`
- » Match time: 1:00 AM, 2:34PM,
- » More difficult match a time after 1pm when expressed using a 24h clock (e.g. a time after 12:00)

PRACTICAL SESSION 2

Write regular expressions which match the given expressions and use your chosen software (Notepad++/Microsoft Word) to perform the replacement

1. You are given a date in the format dd/mm/YYyy convert it to yy-mm-dd (e.g. 11/03/2022 → 22-03-11)

2. Convert numbers from the format XX,XXX.XX to XX.XXX,XX (US format to European)
3. We have a file with a list of terms in English where each term is indicated by the tag <en>. The task is to prepare the file to be translated by duplicating the text, but surrounded by the code of the target language (but not translate the text)

   E.g. `<en>translation memory</en>` →
   `<en>translation memory</en> <ro>translation memory</ro>`

EXTRA EXERCISE

If you have time, try to identify abbreviations and their long form in the following lines using regular expressions. The expressions you should match are underlined below. We will try to cover this exercise if we have time.

Tip: Implement this in regexr.com and to make your life easier do not perform case sensitive matching. (actually I am not sure how you could do it otherwise 😊)

```
Many of the genes are regulated by the transcription factor sterol regulatory element
binding protein (SREBP), which

Gestational diabetes (GDM) is a transient disturbance of glucose metabolism, with a
prevalence ranging from 1.1% to 24.3%, depending on diagnostic criteria [1].

conducted the False Discovery Rate (FDR) [24] method for

are in general unknown anti-retroviral therapy (ART), immune hyper-activation
```

AFTER THE WORKSHOP

Please provide us with feedback using the following link: https://forms.office.com/r/xCWXgfBVrk

The form is anonymous and will open from 13:30 BST. It will be available till Tue, 29th March lunch time.

REGULAR EXPRESSIONS CHEAT SHEET

Expression	Explanation
Abc	Matches the exact string Abc in this case. In most cases the case matters
.	Any character except new line (\n)
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
(a|b)	a or b
(...)	Group
\x	Group/subpattern number "x"
a*	0 or more a (e.g. “”, “a”, “aa”, “aaa”, …)
a+	1 or more (e.g. “a”, “aa”, “aaa”, …)
a?	0 or 1 a (e.g. “”, “a”) (or to make a match not greedy)
a{3}	Exactly 3 a
a{3,}	3 or more a
a{3,5}	3, 4 or 5 a (e.g. “aaa”, “aaaa”, “aaaaa”
^	Start of string, or start of line in multi-line pattern
$	End of string, or end of line in multi-line pattern
\b	Word boundary
\B	Not word boundary
\s	White space
\S	Not white space
\d	Digit
\D	Not digit
\w	Word
\W	Not word