Chapter 06 – Regular Expressions

What's a Regex?

- Most strings like "Hello" are hard coded
- This means, "Hello" means "Hello" and nothing else
- In programming and life, frequently we need to find or match strings that are 'sort of'
- For Example, I want to check if a variable has the string "Hello" or "Jello"
- Maybe if (var == "Hello" || var == "Jello")
- This gets tedious really fast
- Wouldn't it be nice, if we could write a pattern that would mean both Hello and Jello?!

What's a Regex?

- Regex is a pattern instead of a hardcoded string
- For example the regex .ello will match Hello, Jello, Nello, Kello, Cello, Tello… etc

Where to use Regex

- Most programming languages (Python, Java, etc.) allow using regular expressions when comparing strings
- Many utilities that have a search or replace feature let you use regex
    - grep
    - vim
    - awk
    - sed

How to use Regex

- Enabling regex depends on which language or utility you are using
- We shall use grep for practice
- In grep, simply put the regex pattern to search
- Make sure to use the -E option so grep understands extended regex

grep -E ‘regex_search_pattern’ inputfile

Thinking in Regex

- Coming up with regex is like solving puzzle
- Even if you have a ‘correct’ regex you could do better
- Think carefully about what you want to search
- This should be your thought process (for example) – I need my term to start with this char, followed by these chars, followed by one or more of these chars… so on
- Don’t forget to think about what things you DON’T want included in the search!!

Groups and Ranges

.	Any character except new line (\n)
(abc|def)	abc or def
[abc]	a or b or c
[^abc]	Not (a or b or c) = Neither a nor b nor c
[a-q]	Lowercase letter from a to q
[A-Q]	Uppercase letter from A to Q
[0-7]	Digit from 0 to 7

Note: Each of these represents one character where ever you put it in your pattern

Groups and Ranges

- Suppose you want to match Hello and Jello and nothing else

  So you want an H or a J followed by ell

  [HJ]ello

- Suppose you want to match anything that is of the form Hello### where # is a digit

  So you want Hello followed by 3 digits

  Hello[0-9][0-9][0-9]

- Suppose you want to match any string that starts with a digit, followed by ello, followed by a char

  [0-9]ello.

Character Classes

[:upper:]	Upper case letters
[:lower:]	Lower case letters
[:alpha:]	All letters
[:alnum:]	Digits and letters
[:digit:]	Digits
[:xdigit:]	Hexadecimal digits
[:punct:]	Punctuation
[:blank:]	Space and tab
[:space:]	Blank characters

Note: for grep use within [ ], e.g.
[[:alpha:]]

Character Classes

- Suppose you want to match anything that is of the form Hello### where # is a digit

  So you want Hello followed by 3 digits

  Hello[[:digit:]][[:digit:]][[:digit:]]

- Suppose you want to match any string that starts with a digit, followed by ello, followed by a lowercase letter

  [[:digit:]]ello[[:lower:]]

Quantifiers

(Put after a char class/range)

*	0 or more
+	1 or more
?	0 or 1
{3}	Exactly 3
{3,}	3 or more
{3,5}	3, 4 or 5

Quantifiers

- Suppose you want to match anything that is of the form Hello### where # is a digit

  So you want Hello followed by 3 digits

  Hello[[:digit:]]{3}

- Suppose you want to match Hello, Heello, Heeello, basically H followed by 1 or more e's, followed by lo

  He+llo

Anchors

^	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
\<	Start of word
\>	End of word

Anchors

- Suppose you want to match stand-alone Hello (not for example Hellor or THello)

  So word boundary, Hello, word boundary

  \bHello\b

- Suppose you want to match Hello, Heello, Heeello, basically H followed by 1 or more e's, followed by lo, but only if it is at the end of a line

  He+llo$

Assertions

?=	Lookahead assertion
?!	Negative lookahead
?<=	Lookbehind assertion
?!= or ?<!	Negative lookbehind
?()	Condition [if then]
?()|	Condition [if then else]

Practice and Examples

- Write regex for the following:
- Match all email addresses
- Match all phone numbers
- Match website urls
- Don’t match sting or ring, but match all other words that rhyme with bring

Examples / Practice