A Primer on Regular Expression!

Some Points to Ponder!

- A science is any discipline in which the fool of this generation can go beyond the point reached by the genius of the last generation.-Max Gluckman, a South African Antrhopologist
- Science is what we understand well enough to explain to a computer – Don Knuth
- “WORK EXPANDS SO AS TO FILL THE TIME AVAILABLE FOR ITS COMPLETION” - Prof. Cyril Northcote Parkinson (Parkinson’s Law)

What is a Regular Expression?

- A regular expression is a pattern that defines a string or portion thereof. When comparing this pattern against a string, it'll either match or won’t match. It is possible to retrieve partial matches also. (Text Editors , a case in the point)

Regular Sets

Family of languages

- Seed elements:
  - Empty language
  - Language containing the empty string
  - Singleton language for each letter in the alphabet
- Closure Operations:
  - Union: collects strings from languages
  - Concatenation: generates longer strings
  - Kleene Star: generates infinite languages

Regular Set over Sigma

- **Basis**: $\phi, \{\lambda\}$, and $\forall a \in \Sigma : \{a\}$ are regular sets over Σ.
- **Inductive Step**: Let X and Y be regular sets over Σ. Then so are:

$$X \cup Y$$

$$XY$$

$$X^*$$

- **Closure**:...

Regular Expression Basics

- An Empty string matches itself (“”)
- Phi (Null) matches itself
- Any Ascii character which is not a special character matches itself.
  - – A matches A
  - – b matches b
- Concatenation , Alteration and Kleene Closure can be used to create Regular expressions that match complicated Lexemes.

Closure and Regular Expression

Re(NULL) => NULL

Re(“”) => “”

Re([a-z]) => [a-z]

Re.Re => Re

(Re | Re) => Re

Re* => Re

The above stuff defines Re (Recursive definition)

What about R+?

Re+ = Re.Re*

Regular Expression support

- Unix Lex and GNU Flex
- Grep utility
- AWK , SED and Perl
- JavaScript
- C# . C++,Java
- Python,Ruby
- Who does not support it?

Regular Expressions:Special Characters

- A period (.) matches any single character
- A pipe (|) means either what comes before it or what comes after it.
- A caret (^) at the beginning of a RegEx means that the regex will only match if it starts at the beginning of the comparison string
- A dollar sign ($) at the end of a RegEx means that the regex will only match if it ends at the end of the comparison string
- A backslash (\) means escape the next character if it is a special one
- If the character after the backslash is not a special one, then it may be an escape sequence
- Displaying a backslash (\) is done by escaping it

Regular Expressions:Sets

- A character set is a group of characters from which only one is desired [0123456789]
- matches any single number Sets can use ranges of characters (think ascii table) [0-9]
- matches any single character A dash can be represented in a set by placing it first (I.e. not in a range) [-aeiou]
- matches a dash or a vowel A Carat (^) at the beginning of a set negates if (I.e. anything BUT characters in the set

Regular Expressions:Groups

- A group allows a portion of a regular expression to be separated from another portion
- Also known as subexpressions
- Uses parenthesis to group things together

  REFindNoCase(‘(this|that):’, ‘find this:’) = 6

Regular Expressions:Modifiers

- A modifier will take the previous character, set or group and say how many times it can or should exits.
  - REFindNoCase(‘ha+’, ‘hahaha’) = 1
  - REFindNoCase(‘ha*’, ‘hhaha’) = 1
  - REFindNoCase(‘ha?’, ‘hahaha’) = 1
  - REFindNoCase(‘ha{2}’, ‘hahaaha’) = 3
  - REFindNoCase(‘ha{2,3}’, ‘hahaha’) = 3
  - REFindNoCase(‘ha+{3,}’, ‘hahaha’) = 0
  - REFindNoCase(‘(ha)+’, ‘hahaha’) = 1

Regular Expressions (Egs)

- Recognize a Floating-point number
  - (([0-9]+)?\.([0-9]+)?(E|e)?(\+|\-)?[09]+|[0-9]+)|[0-9]+
- Telephone #
  - (([0-9]{2}\-)?([0-9]{3}\-))?[0-9]{7}
- Some Web sites
  - htp://(([wW]{3})?\.)?[a-zA-Z_09]+\.((com|edu|gov)|(co|gov)\.(uk|se))

Q&A

- If any!