RegEx

LIGHTNING TALK 2017-09-15

BY ARNOLD BARNA

Regex (= regular expressions)

- **Text/string processing**; anywhere, where input is textual
  - Validate text
  - Analyse text
  - Extract from text
  - Edit text
  - Replace in text
  - Delete substrings text

Application examples

Data validation

Is hello@@anywhere;com a valid email address?

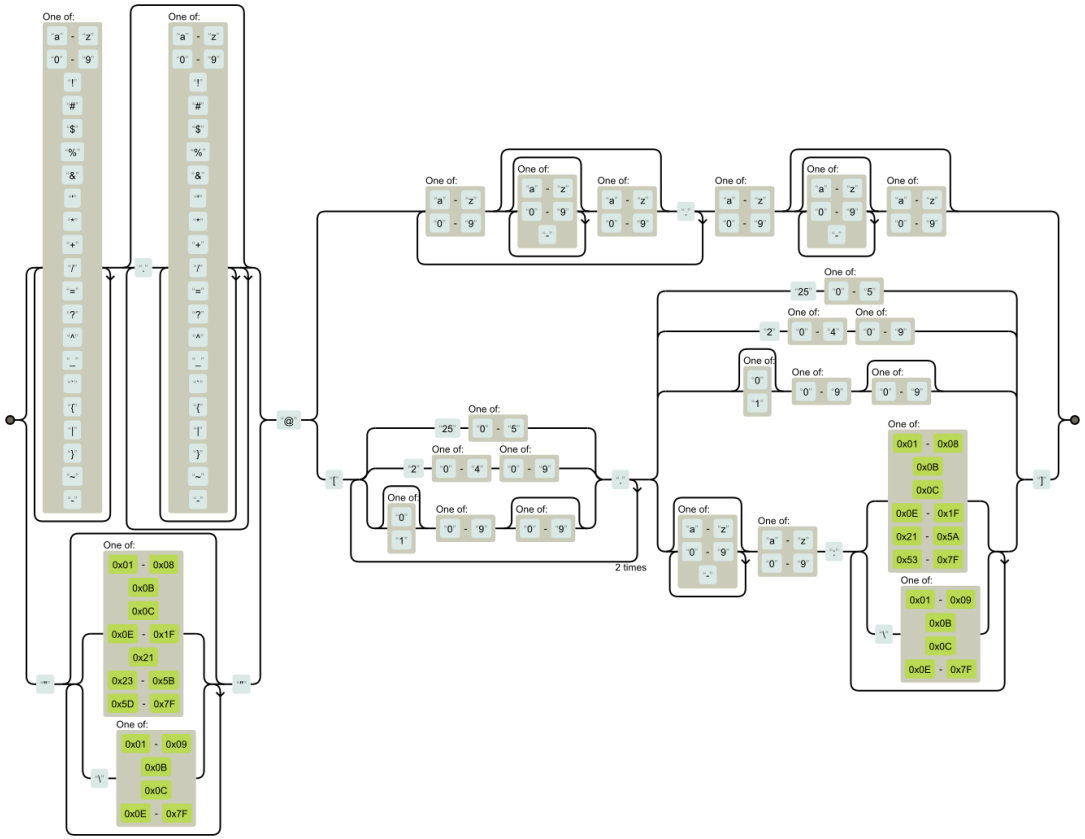
Data/Web scraping

Browser: what is between the tags? ***<head>***^%$@^&#%***</head>***

Data wrangling

(convert from one format to another)

Email validation railroad diagram



Regex

Input for regex processor:

- A string to process
- A string with the **pattern**

General Email Validation Regex

RFC 5322 Official Standard:

- (?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])

.NET (general pattern match)

- ^\w+([-+.']\w+)*@\w+([-.]\w+)*\.\w+([-.]\w+)*$

C#

- ^(?(")(".+?(?<!\\)"@)|(([0-9a-z]((\.(?!\.))|[-!#\$%&'*\+/=\?\^`\{\}\|~\w])*)(?<=[0-9a-z])@))(?(\[)(\[(\d{1,3}\.){3}\d{1,3}\])|(([0-9a-z][-\w]*[0-9a-z]*\.)+[a-z0-9][\-a-z0-9]{0,22}[a-z0-9]))$

Simple Examples

Regex String	Rule	Matches	Doesn’t match
“[abc]”	"a", "b", or "c"	Nebraska	Ohio
“[^hio]”	Anything BUT “h“, “l” or “o”	Ohio Cowboys	Ohio
“v{3}”	“v” three times	Nevvvvada	Nevvada
“yo.”	Any three-character string starting with “yo”	York	Ork
“[^hc]at”	No “h” or “c” before “at”	Rat, Bat	Cat, Hat

Simple Examples

Regex String	Rule	Matches	Doesn't match
"H(ä|ae?)ndel"		"Handel's in da house" "DJ Händel" "Haendel rocks"	"Hndel's works"
\d\d	2 digit	Route 66	5 Stau auf der Autobahn A4
\D\D	2 non-digit	FA155	45823
\w	[A-Za-z0-9_]	4cjv_tr#^$	!@$@

Regex vs. String Methods

Regex: when searching for a **specific pattern** in a string

String Methods: When searching for a **specific string**

Regex vs. String Methods

String Methods	Regex Expression
myString.**StartsWith**(“Green”)	*“**^**Green”*
myString.**EndsWith**(“Fox”)	*“Fox**$**”*
myString.**Contains**(“Alpaga”)	*“Alpaga” or [aglp]*
myString.**Replace**("old", "new")	*“s/old/new/”*

Real world usage

Poker hand and card validation with and without regex

Example from Sortable Poker Hand

French card notation: e.g. *“KS” “2H” “5C”*

- Ranks: 2, 3, 4, 5, 6, 7, 8, 9, T(en), J(ack), Q(ueen), K(ing), A(ce)
- Suits: S(pades), H(earts), D(iamonds), C(lubs)

Valid inputs

Valid card: „<Rank><Suit>”

- Acceptable: *“2H” “td”*
- Not acceptable: *“c9”* “c99” “” etc

Card validation

```
public static bool CardIsValid_WithRegex(string card, char[] Ranks, char[] Suits)
{
    Regex r = new Regex(@"^[23456789TJQKA]{1}[SHDC]{1}$", RegexOptions.IgnoreCase);
    return r.IsMatch(card); ;
}
```

```
public static bool CardIsValid(string card, char[] Ranks, char[] Suits)
{
    card = card.ToUpper();
    return !((card.Length != 2 || !Ranks.Contains(card[0]) || !Suits.Contains(card[1])));
}
```

Card validation pattern

"^[2-9TJQKA][SHDC]$"

^: the string must start with the pattern (String.BeginsWith())

[2-9TJQKA]: Any of these character once (Ranks)

[SHDC]: Any of these character once (Suits)

$: the string must end with the pattern (String.EndsWith()

Thank you for the attention

Available on GitHub:

Code:

- https://github.com/greenfox-academy/**bramble100**/tree/master/week-02/**SortablePokerHands**

Slides:

- https://github.com/greenfox-academy/**bramble100**/tree/master/week-02/**lightning-talk**