Regulární výrazy

Práce s daty, 23. února 2023

Co je to regulární výraz (RE, regexp)?

- Řetězec znaků reprezentující „vyhledávací masku“ v textech
- Kombinuje běžné znaky a zástupné znaky (metaznaky)
  - Písmena, číslice (a další) jsou běžné znaky
  - Tečka je metaznak reprezentující právě jeden libovolný znak
  - B.k je RE odpovídající slovům Bok, Bek, Brk, B7k, B#k, …

„Unix bez regulárních výrazů je jako sex bez partnera/partnerky. Dá se to používat, ale člověk o cosi zásadního přichází. Znalost regulárních výrazů vám dá do rukou mimořádně silný nástroj pro práci s textem.“

- Zdroj: https://www.root.cz/serialy/regularni-vyrazy/

Kde se regulární výrazy používají

- **Validace formulářů**: Kontrola správnosti vstupních údajů, jako je e-mailová adresa, telefonní číslo, datum atd.
- **Textové analytiky**: Vyhledávání a extrakce informací z textových dat, jako jsou např. webové adresy, e-mailové adresy, telefonní čísla atd.
- **Textové úpravy**: Úprava formátování textu, např. odstranění duplicitních řádků nebo bílých znaků.
- **Sémantický parsing**: Rozpoznávání a analýza složených výrazů, jako jsou např. časové značky, čísla, jména či jiné identifikátory atd.
- **Automatický překlad**: Náhrada textových řetězců pomocí regulárních výrazů.
- **Logování**: Analyzování logů a vyhledávání specifických informací.
- **Zpracování dat**: Filtrace a úprava dat pomocí regulárních výrazů.
- **Zabezpečení**: Kontrola nebezpečných vstupů, jako jsou např. SQL injection nebo XSS útoky.

Varianty regulárních výrazů

- Standard IEEE Posix
  - Basic Regular Expressions (BRE)
  - Extended Regular Expressions (ERE)
- Perl Compatible Regular Expressions (PCRE)
  - Vychází z implementace v programovacím jazyce Perl
  - Stal se de-facto standardem v dalších jazycích a systémech
- Rozšíření
  - Podpora národních abeced – Unicode
  - Kondicionály (if – then – else)
  - Vnořené modifikátory, komentáře, pojmenované skupiny a další...

Zástupné znaky v RE

- **.** = libovolný znak
- **\d** = číslice (arabská desítková, tj. v rozsahu 0–9)
- **\w** = znak „slova“ (číslice, písmena z ASCII a podtržítko)
- **\s** = bílé či neviditelné znaky (mezera, tabulátor, konec řádku)
- **\D**, **\W**, **\S** = negace (cokoliv vyjma) \d, \w, \s
- **^** = začátek řetězce (řádku)
- **$** = konec řetězce (řádku)
- **\b** = hranice slova
  - hranice jsou „logické“ indikátory, samy o sobě neobsahují žádný znak

Další speciální znaky v RE

- **\n** = nový řádek (ve Windows je nový řádek = **\n\r**)
- **\t** = tabelátor
- **** = znak za zpětným lomítkem nebude interpretován jako metaznak
  - **\.** = tečka
  - **\^** = stříška
  - **\$** = dolar
  - **** = zpětné lomítko
- **|** = alterace (logická spojka nebo)
- **( )** = seskupení znaků

Výčty a opakování v RE

- **+** = jedno a více opakování
- ***** = žádné a více opakování
- **?** = žádné nebo jedno opakování
- **{*m*}** = právě *m* opakování
- **{*m*,*n*}** = minimálně m opakování, maximálně n opakování
- **{*m*,}** = minimálně m opakování, maximálně neomezeno
- **[ ]** = výčet znaků, např. [x-z], [adfr]
- **[^]** = negace výčtu znaků

Cvičení I

Popište, jakým řetězcům odpovídají následující RE:

1. `.*`
2. `[abc]`
3. `ISSN \d{4}-\d{4}`
4. `^$`
5. `\baby\b`
6. `(ko)+dák`
7. `^[^\t]+\t`

Cvičení I – řešení

Popište, jakým řetězcům odpovídají následující RE:

1. **`.*`** – jakýkoliv řetězec, včetně prázdného
2. **`[abc]`** – písmeno a, b nebo c
3. **`ISSN \d{4}-\d{4}`** – např. ISSN 1578-0026
4. **`^$`** – výhradně prázdný řetězec
5. **`\baby\b`** – slovo „aby“ (nikoliv však „baby“, „abychom“ apod.)
6. **`(ko)+dák`** – kodák, kokodák, kokokodák, kokokokodák atd.
7. **`^[^\t]+\t`** – cokoliv od začátku po první tabelátor

Cvičení II

Zapište RE, které budou validovat následující vstupy

1. **Jméno Příjmení** (bez diakritiky, dvě slova oddělená mezerou)
2. **Internetová doména** (www.econ.muni.cz, blog.google)
3. **Krevní skupina včetně Rh faktoru** (A, B, AB, 0 ve variantách + a -)
4. **PSČ v rámci České republiky** (bez Slovenska)
5. **Sudý počet znaků** (nebo také lichý počet znaků)
6. **Kód barvy v CSS** (hexadecimální notace, např. #45af09 nebo #09f)
7. **IP-adresa** (0.0.0.0–255.255.255.255)

Cvičení II – řešení

Zapište RE, které budou validovat následující vstupy

1. **Jméno Příjmení** – `[A-Z][a-z]+ [A-Z][a-z]+`
2. **Internetová doména** – `[\w\.-]+\.\w{2,}`
3. **Krevní skupina včetně Rh faktoru** – `([AB0]|AB)(\+|-)`
4. **PSČ v rámci České republiky** – `[1-7]\d{2} \d{2}`
5. **Sudý / lichý počet znaků** – ^(..)+$ ^.(..)*$
6. **Kód barvy v CSS** – #?([a-fA-F0-9]){3}(([a-fA-F0-9]){3})?
7. **IP-adresa** – ((25[0-5]|2[0-4]\d|[01]\d\d|\d?\d)\.){3}(25[0-5]|2[0-4]\d|[01]\d\d|\d?\d)

Cvičení (nebo) Domácí úkol (?)

Úspěšně projít první čtyři bloky procvičování RE na

https://www.umimeinformatiku.cz/regularni-vyrazy

(co denní limit dovolí)

Návody na Internetu

- Regulární výraz včetně **formální definice** na Wikipedii
- Základy regulárních výrazů na www.regularnivyrazy.info
- Seriál návodů od Pavla Satrapy
- Tester regulárních výrazů www.regexp.cz/
- Editory regexr.com, regex101.com, regexpal.com
- E-kniha Introducing regular expressions
- Knihovna často používaných RE: www.regexlib.com
- Tak trochu jiný tutoriál k RE: www.rexegg.com