

Předmluva	7
Uvěznění mezi verzemi	7
Cílová skupina	8
Popisované technologie	8
Jak je kniha členěna	8
Rozlišení textu	9
Příklady kódu	10
Safari® Books Online	10
Jak se s námi spojit	11
Poděkování	11
Poznámka redakce českého vydání	11
1. Úvod do regulárních výrazů	13
Definice regulárních výrazů	13
Vyhledávání a nahrazování pomocí regulárních výrazů	17
Nástroje pro práci s regulárními výrazy	18
2. Základy práce s regulárními výrazy	35
2.1 Hledání shodného literálního textu	35
2.2 Netisknutelné znaky	37
2.3 Hledání jednoho znaku mezi mnohými	39
2.4 Hledání libovolného znaku	43
2.5 Vyhledávání na začátku a konci řádku	45
2.6 Hledání celých slov	49
2.7 Body kódu, vlastnosti, bloky a jazyky ve standardu Unicode	51
2.8 Vyhledání jedné z několika možností	61
2.9 Skupinové a zachytávací části shody	63
2.10 Opětovné vyhledání již vyhledaného textu	65
2.11 Zachycení a pojmenování části shody	67
2.12 Několikanásobné opakování části regulárního výrazu	69
2.13 Minimální a maximální počet opakování	72

2.14	Eliminace zbytečného zpětného vyhledávání	74
2.15	Zabraňujeme nechtěnému opakování	77
2.16	Vyhledání shody bez vložení do výsledků vyhledávání	79
2.17	Vyhledání jedné z alternativ v závislosti na podmínce	84
2.18	Vkládání komentářů do regulárního výrazu	86
2.19	Vkládání literálního textu do nahrazovacího textu	88
2.20	Vkládání shody regulárního výrazu do nahrazovacího textu	90
2.21	Vkládáme část shody regulárního výrazu do nahrazovacího textu	91
2.22	Vkládání kontextu shody do nahrazovacího textu	94

3. Programování s regulárními výrazy **97**

	Programovací jazyky a regulární výrazy	97
3.1	Literální regulární výrazy ve zdrojovém kódu	101
3.2	Import knihovny regulárních výrazů	106
3.3	Vytváříme objekty regulárních výrazů	107
3.4	Nastavení voleb regulárních výrazů	113
3.5	Test na výskyt shody ve zdrojovém řetězci	118
3.6	Test úplné shody regulárního výrazu ve zdrojovém řetězci	124
3.7	Načtení shodného textu	128
3.8	Určení pozice a délky shody	133
3.9	Načítání části textu shody	138
3.10	Načtení seznamu všech shod	144
3.11	Iterování na všech shodách	148
3.12	Ověření shody v procedurálním kódu	153
3.13	Hledání shody v jiné shodě	156
3.14	Nahrazení všech shod	160
3.15	Nahrazování shod se znovuvyužitím jejich částí	166
3.16	Nahrazování shody textem vygenerovaným kódem	170
3.17	Nahrazování všech shod jiného regulárního výrazu	175
3.18	Nahrazení všech shod nacházejících se mezi shodami jiného regulárního výrazu	177
3.19	Rozdělení řetězce	182
3.20	Rozdělení řetězce při zachování shody regulárního výrazu	190
3.21	Prohledávání řádku po řádku	193

4. Ověřování a formátování **197**

4.1	Ověřování e-mailových adres	197
4.2	Ověřování a formátování severoamerických telefonních čísel	202
4.3	Ověřování mezinárodních telefonních čísel	206
4.4	Ověřování dat v tradičním formátu	208
4.5	Přesné ověřování dat v tradičních formátech	211
4.6	Ověřování času v tradičních formátech	216
4.7	Ověřování datových a časových údajů ve standardu ISO 8601	218

4.8	Omezení vstupu na alfanumerické znaky	223
4.9	Omezení délky textu	225
4.10	Omezení počtu řádků v textu	229
4.11	Ověřování kladných hodnot v odpovědích	233
4.12	Ověřování čísel sociálního pojištění	234
4.13	Ověřování čísel ISBN	237
4.14	Ověřování amerických poštovních směrovacích čísel	243
4.15	Ověřování kanadských poštovních směrovacích čísel	244
4.16	Ověřování britských poštovních směrovacích čísel	245
4.17	Vyhledávání adres s čísly P.O. Boxů	246
4.18	Převedení jmen z formátu „KřestníJméno Příjmení“ na „Příjmení, KřestníJméno“	247
4.19	Ověřování čísel kreditních karet	250
4.20	Evropská čísla plátců DPH	256

5. Slova, řádky a zvláštní znaky

261

5.1	Vyhledání konkrétního slova	261
5.2	Vyhledání jednoho slov z mnoha	263
5.3	Vyhledávání podobných slov	265
5.4	Vyhledání všeho s výjimkou konkrétního slova	268
5.5	Vyhledávání slova nenásledovaného zadaným slovem	270
5.6	Vyhledávání slov, před nimiž se nenachází definované slovo	271
5.7	Vyhledávání slov nacházejících se nedaleko od sebe	274
5.8	Vyhledávání opakujících se slov	280
5.9	Odstranění duplicitních řádků	281
5.10	Vyhledávání celých řádků obsahujících požadované slovo	285
5.11	Vyhledávání celých řádků neobsahujících dané slovo	286
5.12	Ořezání počátečních a koncových bílých mezer	287
5.13	Nahrazování opakujících se bílých mezer jedinou mezerou	290
5.14	Uvozování metaznaků	291

6. Čísla

295

6.1	Celá čísla	295
6.2	Šestnáctková čísla	298
6.3	Binární čísla	300
6.4	Ořezání počátečních nul	301
6.5	Čísla v daném rozsahu	302
6.6	Šestnáctková čísla v daném rozsahu	308
6.7	Čísla s plovoucí desetinnou tečkou	311
6.8	Čísla s oddělovači tisíců	313
6.9	Římské číslice	314

7. URL, cesty a internetové adresy

317

7.1	Ověřování adres URL	317
7.2	Fulltextové vyhledávání adres URL	320
7.3	Fulltextové vyhledávání citovaných adres URL	321
7.4	Fulltextové vyhledávání adres URL obsahujících závorky	322
7.5	Převádění adres URL na odkazy	324
7.6	Ověřování identifikátorů URN	325
7.7	Ověřování generických adres URL	327
7.8	Načítání vzoru z adresy URL	332
7.9	Zjišťování jména uživatele z adresy URL	334
7.10	Zjišťování hostitele z adresy URL	335
7.11	Zjišťování čísla portu z adresy URL	337
7.12	Zjišťování cesty z adresy URL	339
7.13	Načítání požadavku z adresy URL	342
7.14	Načítání fragmentu z adresy URL	343
7.15	Ověřování názvů domén	343
7.16	Vyhledávání adres IPv4	345
7.17	Vyhledávání IPv6 adres	348
7.18	Ověřování cesty v systému Windows	361
7.19	Rozdělení cesty v systému Windows na části	363
7.20	Zjišťování označení diskového oddílu z cesty v systému Windows	368
7.21	Zjišťování názvu serveru a sdíleného umístění ze síťové cesty	369
7.22	Zjišťování adresáře z cesty v systému Windows	370
7.23	Zjišťování názvu souboru z cesty v systému Windows	372
7.24	Zjišťování přípony souboru z cesty v systému Windows	373
7.25	Jak z názvů souborů odstranit neplatné znaky	373

Rejstřík

375

Předmluva

Popularita regulárních výrazů za poslední desetiletí značně narostla. Všechny programovací jazyky dnes obsahují knihovnu výkonných regulárních výrazů, anebo mají jejich podporu přímo zabudovanou. Mnozí vývojáři funkcí těchto regulárních výrazů využívají a uživatelům svých aplikací jejich prostřednictvím umožňují data vyhledávat a filtrovat. Regulární výrazy jsou všudypřítomné.

Na vlně nasazování regulárních výrazů se svezlo již mnoho knih. Většinou z nich se daří vysvětlit syntaxi regulárních výrazů, jejich příklady a reference dobře. Nenačtete však knihu, která by na základě regulárních výrazů nabízela řešení široké škály praktických problémů z reálného života, s nimiž se setkáváme při práci s textem na počítači a v množství internetových aplikací. Jako autoři knihy (Steve a Jan) jsme se rozhodli, že zaplníme právě tuto mezeru.

Obzvláště vám chceme ukázat, jak můžete regulární výrazy užívat v situacích, ve kterých by vám lidé s omezenými zkušenostmi v práci s regulárními výrazy řekli, že je problém neřešitelný, anebo v takových situacích, v nichž by softwaroví puristi tvrdili, že se regulární výrazy na danou problematiku nehodí. Protože jsou dnes regulární výrazy všude okolo, mohou s nimi začít koncoví uživatelé přímo pracovat, aniž by museli do práce zapojovat tým programátorů. Dokonce i programátoři mohou nasazením regulárních výrazů ušetřit čas při získávání dat a změně úloh, jejichž ošetření by v procedurálním kódu trvalo hodiny či dny, nebo jejichž použití by vyžadovalo knihovnu třetí strany, kterou je třeba zkontrolovat a nechat schválit managementem.

Uvěznění mezi verzemi

Stejně jako je to se vším, co se stává v průmyslové oblasti výpočetních technologií populárním, i regulární výrazy přichází v mnoha různých implementacích, které se liší stupněm kompatibility. Vznikla tak řada různých *příchutí* regulárních výrazů. Ty s daným regulárním výrazem nezachází vždy stejně, případně s ním nezachází vůbec nijak.

Mnoho knih tyto různé příchuti zmiňuje a vyzdvihuje i některé z rozdílů, které se mezi nimi nachází. Obvykle však některé z příchutí přeskakují – zvláště tehdy, když jim chybí některé konkrétní funkce – namísto toho, aby nabídly alternativní řešení či alespoň způsob, jak nedostatek obejít. Když pak máte pracovat s různými druhy regulárních výrazů ve více aplikacích či programovacích jazycích, budí to ve vás úzkost.

Ležerní výroky v literatuře ve stylu „dnes všichni pracují s perlovskými regulárními výrazy“ bohužel bagatelizují celou škálu problémů týkajících se nekompatibility. Dokonce i mezi „perlovskými“ balíky jsou význačné rozdíly, a Perl se přitom i nadále vyvíjí. Přesvědčíš zjednodušené výrazy mohou způsobit, že programátor stráví kupříkladu půl hodiny bezcílným laděním, namísto toho, aby zkontroloval podrobnosti týkající se své implementace regulárního výrazu. I když pak zjistíš, že některá z funkcí, na kterou se spoléhal, není dostupná, neví, jak si s problémem poradit.

Tato kniha je první publikací na trhu, která v celém svém průběhu celistvě srovnává nejoblíbenější a na funkce nejbohatší implementace regulárních výrazů.

Cílová skupina

Jestli na počítači pracujete pravidelně s texty, měli byste si tuto knihu přečíst. Nezáleží na tom, zda se prohrabáváte hromadou dokumentů, pracujete s textem v textovém editoru, anebo vyvíjíte software, který v textu musí umět vyhledávat či s ním manipulovat. V těchto případech jsou regulární výrazy fantastickým nástrojem. *Regulární výrazy Kuchařka programátora* vás naučí vše, co o regulárních výrazech potřebujete vědět. Nemusíte mít žádné předešlé zkušenosti, protože vám vysvětlíme i ty nejzákladnější aspekty regulárních výrazů.

Jestli ještě s regulárními výrazy nemáte zkušenosti, naleznete v této knize studnici podrobností, které jiné knihy a články na internetu často přehlíží. Jestliže jste se již někdy dostali do situace, kdy regulární výraz v jedné aplikaci fungoval, ale v jiné ne, bude se vám podrobný a rovnocenný popis sedmi světově nejoblíbenějších regulárních výrazů velice hodit. Celou knihu jsme uspořádali do podoby kuchařky, takže můžete přeskakovat přímo na témata, o kterých se chcete něco dočíst. Budete-li ji číst od začátku do konce, stane se z vás prvotřídní šéfkuchař regulárních výrazů.

Tato kniha vás o regulárních výrazech naučí vše, co potřebujete znát, a ještě i něco navíc. Nezáleží na tom, zda jste programátor nebo ne. Chcete-li regulární výrazy použít při práci s textovým editorem, vyhledávacím nástrojem či jinou aplikací, která vyžaduje zadání regulárního výrazu, nemusíte mít k četbě této knihy s regulárními výrazy žádné zkušenosti. Většina z receptů v této knize zakládá řešení čistě na jednom či několika regulárních výrazech.

Jste-li programátor, 3. kapitola vám poskytne veškeré informace potřebné k implementaci regulárních výrazů ve zdrojovém kódu. Kapitola předpokládá, že znáte základní programovací funkce vámi vybraného programovacího jazyka, ale nepočítá s tím, že byste již někdy ve zdrojovém kódu regulární výrazy používali.

Popisované technologie

Hesla .NET, Java, JavaScript, PCRE, Perl, Python a Ruby nejsou na zadní straně přebalu uvedeny jen ze zvyku. Jedná se o oněch sedm příchutí regulárních výrazů, kterým se v knize budeme věnovat. Všemi se přitom zabýváme se stejnou intenzitou. Zvláště se snažíme popsat všechny nesoulady, které by se mezi těmito příchutěmi mohly vyskytnout.

Kapitola pro programátory (Kapitola 3) pracuje s výpisy kódu v jazycích C#, Java, JavaScript, PHP, Perl, Python, Ruby a VB.NET. I zde je v každém receptu zahrnuto řešení a vysvětlení všech osmi jazyků. I když se tak v kapitole trochu opakujeme, můžete diskuze o jednotlivých jazycích, které vás nezajímají, jednoduše přeskočit, aniž byste přišli o něco, co byste o vámi vybraném jazyce měli vědět.

Jak je kniha členěna

První tři kapitoly knihy se věnují užitečným nástrojům a základním informacím, které vám poskytnou základy používání regulárních výrazů. Všechny podřízené kapitoly představují škálu regulárních výrazů, přičemž se vždy hlouběji věnují jedné oblasti zpracování textu.

Kapitola 1, *Úvod do regulárních výrazů*, osvětluje roli regulárních výrazů a uvádí množství nástrojů, s nimiž se výrazy snáze učí, vytváří a ladí.

Kapitola 2, *Základy při práci s regulárními výrazy*, se věnuje všem prvkům a funkcím regulárních výrazů, spolu s důležitými návody pro jejich efektivní použití.

Kapitola 3, *Programování s regulárními výrazy*, specifikuje techniky kódu a zahrnuje výpisy kódu pro práci s regulárními výrazy ve všech programovacích jazycích, jimiž se v knize zaobíráme.

Kapitola 4, *Ověřování a formátování*, obsahuje recepty na správu s běžnými uživatelskými vstupy, například daty, telefonními čísly a poštovními směrovacími čísly v různých státech.

Kapitola 5, *Slova, řádky a zvláštní znaky*, zkoumá nejběžnější úlohy při práci s textem, například hledání řádků obsahujících či neobsahujících konkrétní slova.

Kapitola 6, *Čísla*, ukazuje, jak lze detekovat celá čísla, čísla s plovoucí desetinnou čárkou a některé další formáty tohoto vstupního druhu.

Kapitola 7, *URL, cesty a internetové adresy*, vám ukáže, jak rozebrat a použít řetězce, se kterými se běžně pracuje na internetu a na systémech Windows při vyhledávání.

Rozlišení textu

V knize pracujeme v souladu s následujícími typografickými konvencemi:

Kurzíva

Značí nové výrazy, adresy URL, e-mailové adresy, názvy a přípony souborů.

Písmo s konstantní šířkou znaků

Toto písmo používáme ve výpisech programů, programových prvcích, jakými jsou například proměnné či názvy funkcí, hodnot vrácených coby výsledky náhrad regulárních výrazů, a v zdrojovém či vstupním textu aplikovaném v regulárním výrazu. Může se jednat o obsah textového pole v aplikaci, souboru na disku, nebo obsahu řetězcové proměnné.

Kurzíva s konstantní šířkou znaků

Písmo používáme k zobrazení textu, který by měl uživatel nahradit svými vlastními hodnotami, nebo hodnotami vycházejícími z kontextu.

<Regulární●výraz>

Představuje regulární výraz stojící samostatně, anebo tak, jak byste ho zadávali do vyhledávacího pole aplikace. Mezery jsou v regulárních výrazech, s výjimkou případů, kdy se s nimi pracuje v režimu free-spacing (ignorování bílých znaků), značeny černými tečkami.

«Náhradní●text»

Představuje text, kterým se v rámci hledacích a nahrazujících operací nahradí shodné regulární výrazy. Mezery jsou v náhradních textech značeny černou tečkou.

Shodný text

Představuje část zdrojového textu, který je shodný s regulárním výrazem.

...

Šedá elipsa v regulárním výrazu značí to, že je třeba „vypsat prázdné pole“ dříve, než budete moci použít regulární výraz. Doprovodný text vysvětluje, co můžete do pole zadat.

`CR`, `LF` a `CRLF`

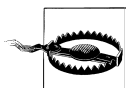
`CR`, `LF` a `CRLF` v rámečcích představují v řetězci vlastní znaky pro zalomení řádku namísto znaků `\r`, `\n` a `\r\n`. Tyto řetězce vznikají v aplikaci třeba ve víceřádkovém editoru stisknutím klávesy `Enter`, nebo při práci s víceřádkovými řetězcovými konstantami ve zdrojovém kódu, například u doslovných řetězců, či v Pythonu u řetězců s trojitými uvozovkami.

↵

Zpětná šipka, jakou vidíte na klávesnici na tlačítku `Enter`, značí, že bylo třeba řádek zalomit, aby se vešel na šířku zobrazené stránky. Při zadávání textu do zdrojového kódu byste neměli používat klávesu `Enter`, ale naopak byste měli všechno zapisovat na jeden řádek.



Tato ikona značí tip, návrh či obecnou poznámku.



Tato ikona předznamenává varování či upozornění.

Příklady kódu

Tato kniha by vám měla pomoci při práci. Kód uvedený v knize můžete ve svých programech a aplikacích volně používat. Jestliže nebudete kopírovat velkou část kódu, nemusíte nás kontaktovat a žádat o svolení. Například k programu, ve kterém použijete několik úryvků kódu z této knihy, naše svolení potřebovat nebudete. K prodeji či redistribuci CD s příklady z knih nakladatelství O'Reilly byste již svolení potřebovali. Zodpovězení otázky citováním této knihy a příkladem kódu si svolení vlastníka práv nežádá. K zařazení většího množství vzorového kódu obsaženého v této knize do dokumentace produktu je však již svolení třeba.

Jakkoliv to nevyžadujeme, byli bychom rádi, kdybyste nám za naši práci přiznali zásluhy. V anotaci se obvykle objevuje titul, autor, nakladatel a číslo ISBN. Například „Regulární výrazy Kuchařka programátora od Jana Goyvaertse a Stevena Levithana. Copyright 2009 Jan Goyvaerts a Steven Levithan, 978-0-596-2068-7.“.

Máte-li pocit, že se chystáte příklady kódu použít nad rámec zde povolených situací, neváhejte nám napsat na adresu permissions@oreilly.com.

Safari® Books Online



Když na přebalu své oblíbené technické knihy uvidíte ikonu Safari® Books Online, znamená to, že knihu lze dohledat i v knihovně O'Reilly Network Safari Bookshelf.

Safari nabízí lepší řešení než elektronické knihy. Poskytuje virtuální knihovnu, s níž budete moci snadno prohledávat tisíce nejlepších technicky zaměřených knih, kopírovat vzorky kódu, stahovat kapitoly a hledat rychle odpovědi, když potřebujete nejpřesnější a nejaktuálnější informace. Službu si můžete zdarma vyzkoušet na adrese <http://my.safaribooksonline.com>.

Jak se s námi spojit

Komentáře a otázky týkající se této knihy prosím směřujte na nakladatele:

O'Reilly Media, Inc.

1005 Gravenstein Highway North

Sebastopol, CA 95472

800-998-9938 (ve Spojených státech amerických či v Kanadě)

707-829-0515 (mezinárodní nebo místní)

707-829-0104 (fax)

K této knize provozujeme i webovou stránku, na které uvádíme seznam chyb, příkladů a další informace. Stránku naleznete na adrese:

<http://www.regexcookbook.com>

nebo na adrese:

<http://oreilly.com/catalog/9780596520687>

Chcete-li knihu komentovat, anebo položit technický dotaz, který se jí týká, zašlete e-mail na adresu:

bookquestions@oreilly.com

Další informace o našich knihách, konferencích, vzdělávacích zařízeních a službě O'Reilly Network hledejte na adrese:

<http://www.oreilly.com>

Poděkování

Děkujeme Andymu Oramovi, našemu redaktorovi ve společnosti O'Reilly Media, Inc, za pomoc při realizaci celého projektu. Za důkladné překontrolování technické stránky knihy patří naše díky také Jeffreymu Friedlovi, Zaku Greantovi, Nikolaji Lindbergovi a Ianu Morseovi. Kniha je díky nim srozumitelnější a přesnější.

Poznámka redakce českého vydání

I nakladatelství Computer Press, které pro vás tuto knihu přeložilo, stojí o zpětnou vazbu a bude na vaše podněty a dotazy reagovat. Můžete se obrátit na následující adresy:

Computer Press
redakce PC literatury
Holandská 8
639 00 Brno

nebo

knihy@cpress.cz

Další informace a případné opravy českého vydání knihy najdete na internetové adrese <http://knihy.cpress.cz/K1577>. Prostřednictvím uvedené adresy můžete též naši redakci zaslat komentář nebo dotaz týkající se knihy. Na vaše reakce se srdečně těšíme.