



به نام خدا

وب کاوی با

Scrapy, Selenium, Beautiful soup

مؤلفان:

نیما شفیعی رضوانی نژاد
افشین اسمعیل زاد آهندانی



مؤسسه فرهنگی هنری
دیباجران تهران

هر گونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.

◀ عنوان کتاب: **وب کاوی با**

Scrapy, Selenium, Beautiful soup

◀ مولفان: **نیما شفیعی رضوانی نژاد - افشین اسمعیل زاد آهندانی**

◀ ناشر: **موسسه فرهنگی هنری دیباجران تهران**

◀ **ویراستار: مهدیه مخبری**

◀ **صفحه آرای: فرنوش عبدالهی**

◀ **طراح جلد: محمد اسمعیل زاد آهندانی**

◀ **نوبت چاپ: اول**

◀ **تاریخ نشر: ۱۴۰۱**

◀ **چاپ و صحافی: صدف**

◀ **تیراژ: ۱۰۰ جلد**

◀ **قیمت: ۱۲۰۰۰۰۰ ریال**

◀ **شابک: ۹۷۸-۶۲۲-۲۱۸-۶۴۹-۴**

◀ **نشانی واحد فروش: تهران، خیابان انقلاب، خیابان دانشگاه**

◀ **-تقاطع شهدای ژاندارمری - پلاک ۱۵۸ ساختمان دانشگاه -**

◀ **طبقه دوم - واحد ۴ تلفن ها: ۶۶۹۶۵۷۴۹-۲۲۰۸۵۱۱۱**

◀ **فروشگاههای اینترنتی دیباجران تهران :**

WWW.MFTBOOK.IR

www.dibagaran-tehran.com

سرشناسه: شفیعی رضوانی نژاد، نیما، ۱۳۷۱-

عنوان و نام پدیدآور: وب کاوی با

scrapy, beautifulsoup, selenium / مولفان: نیما شفیعی

رضوانی نژاد، افشین اسمعیل زاد آهندانی؛

ویراستار: مهدیه مخبری .

مشخصات نشر: تهران: دیباجران تهران: ۱۴۰۱

مشخصات ظاهری: ۱۵۲ ص: مصور، جدول

شابک: ۹۷۸-۶۲۲-۲۱۸-۶۴۹-۴

وضعیت فهرست نویسی: فیبا

موضوع: داده کاوی data mining

موضوع: اسکریپی (software framework) scrapy

موضوع: داده های کلان big data

شناسه افزوده: اسمعیل زاد آهندانی، افشین، ۱۳۷۱-

رده بندی کنگره: ۷۶/۹ QA

رده بندی دیویی: ۰۰۶/۳۱۳

شماره کتابشناسی ملی: ۹۱۰۶۸۰۸

نشانی اینستاگرام دیبا dibagaran_publishing نشانی تلگرام: @mftbook

هر کتاب دیباجران، یک فرصت جدید علمی و شغلی.

هر گوشی همراه، یک فروشگاه کتاب دیباجران تهران.

از طریق سایتهای دیباجران، در هر جای ایران به کتابهای ما دسترسی دارید.

فهرست مطالب

فصل ۱ / معرفی اسکریپت ۹

- ۱۰ سلام اسکریپت
- ۱۱ دلایل بیشتر برای مطلوبیت اسکریپت
- ۱۳ توسعه برنامه‌های قوی و با کیفیت و ارائه زمانبندی‌های واقعی
- ۱۴ توسعه سریع حداقل محصولات قابل عرضه با کیفیت
- ۱۶ کشف و ادغام در اکوسیستم شما
- ۱۶ شهروند خوبی بودن در دنیایی پر از عنکبوت‌ها
- ۱۸ آنچه اسکریپت نیست
- ۱۹ خلاصه

فصل ۲ / آشنایی با HTML و XPATH ۲۰

- ۲۱ HTML، بازنمایی درختی DOM، و XPATH
- ۲۲ URL
- ۲۲ سند HTML
- ۲۴ بازنمایی درختی
- ۲۵ آنچه بر روی صفحه می‌بینید
- ۲۶ انتخاب عناصر HTML با XPATH
- ۲۷ عبارات مفید XPATH
- ۳۰ استفاده از CHROME برای دستیابی به عبارات XPATH

۳۰ نمونه‌هایی از وظایف رایج
۳۲ پیش‌بینی تغییرات
۳۳ خلاصه

فصل ۳ / کراولینگ پایه ۳۴

۳۶ نصب اسکریپی
۳۶ سیستم‌عامل MAC
۳۷ WINDOWS
۳۷ توزیع UBUNTU یا DEBIAN لینوکس
۳۸ توزیع RED HAT یا CENTOS لینوکس
۳۸ ارتقاء اسکریپی
۴۱ UR ² IM - فرایند پایه‌ای اسکریپینگ
۴۱ URL
۴۳ درخواست و پاسخ
۴۴ آیتم‌ها
۴۹ یک پروژه اسکریپی
۵۰ تعریف آیتم‌ها
۵۲ نوشتن عنکبوت
۵۶ پرکردن یک آیتم
۵۷ ذخیره در فایل‌ها
۶۰ پاکسازی - بارگذارهای آیتم و فیلدهای پیشینه
۶۴ ایجاد قراردادها

۶۷.....	استخراج URL های بیشتر
۷۰	کراولینگ دو جهته با استفاده از یک عنکبوت
۷۳	کراولینگ دو جهته با CRAWLSPIDER
۷۵	خلاصه

فصل ۴ / دستور العمل‌های عنکبوت سریع ۷۶

۷۷	عنکبوتی که لاگین می کند
۸۳	یک عنکبوت که از API های JSON و صفحات AJAX استفاده می کند
۸۶.....	انتقال آرگومان‌ها در میان پاسخ‌ها
۸۷	یک عنکبوت ملک با سرعت ۳۰ برابر
۹۲	عنکبوتی که کراولینگ را براساس یک فایل اکسل انجام می دهد
۹۶.....	خلاصه

فصل ۵ / SELENIUM ۹۷

۱۰۱.....	تفکر دربارهٔ مدل‌های خزندهٔ وب
۱۰۲.....	اسکرپینگ با سلنیوم
۱۰۳.....	ظهور سلنیوم ۲
۱۰۳.....	معرفی سلنیوم IDE
۱۰۴.....	مشخصه‌ها

فصل ۶ / BEAUTIFULSOUP ۱۲۲

۱۲۳.....	نصب BEAUTIFULSOUP
۱۲۵.....	اجرای BEAUTIFULSOUP
۱۲۸.....	تجزیهٔ HTML پیشرفته

۱۳۳..... BEAUTIFULSOUP اشیاء

۱۳۷..... BEAUTIFUL SOUP عبارات منظم و

۱۳۸..... دسترسی به ویژگی‌ها

۱۴۱..... WEB SCRAPING پیاده‌سازی فصل ۷

۱۴۲..... ایجاد چند پروژه

۱۴۲..... REQUESTS استفاده از کتابخانه

۱۴۷..... PANDAS خواندن جداول در وب با استفاده از کتابخانه

۱۴۹..... خواندن محتوای وب پویا با سلنیوم

خط‌مشی انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌هایی با کیفیت عالی است که بتواند
خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.
هر کتاب دیباگران تهران، یک فرصت جدید شغلی و علمی

حمد و سپاس ایزد منان را که با الطاف بی‌کران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی و آموزشی گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌گی علوم و سرعت توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه، نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش برجسته نموده است.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران و محققان در زمینه‌های گوناگون و مورد نیاز جامعه تلاش نموده برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهد.

کتابی که در دست‌دارید تألیف "جناب آقایان: نیما شفیعی رضوانی نژاد - افشین اسمعیل زاد آهندانی" است که با تلاش همکاران ما در نشر دیباگران تهران منتشر گشته و شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

با نظرات خود مشوق و راهنمای ما باشید

با ارائه نظرات و پیشنهادات و خواسته‌های خود، به ما کمک کنید تا بهتر و دقیق‌تر در جهت رفع نیازهای علمی و آموزشی کشورمان قدم برداریم. برای رساندن پیام‌هایتان به ما از رسانه‌های دیباگران تهران شامل سایتهای فروشگاهی و صفحه اینستاگرام و شماره‌های تماس که در صفحه شناسنامه کتاب آمده استفاده نمایید.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
dibagaran@mftplus.com

مقدمه مولف

در سال های اخیر حجم داده های متنی که به وسیله وب در دسترس عموم قرار گرفته به طور باورنکردنی افزایش یافته و تحلیل، بررسی و واکاوی این داده ها تا حد زیادی از عهده انسان خارج شده و نیاز به یک نرم افزاری که بسیاری از کارهای تکراری و طاقت فرسای جمع آوری و تحلیل داده ها از وب را انجام بدهد بسیار احساس می شود. توانایی نوشتن یک ربات ساده که داده ها را جمع آوری می کند و آن ها را در یک ترمینال پخش می کند یا آن ها را در یک پایگاه داده ذخیره می کند، اگرچه دشوار نیست، اما همیشه و حس خاصی را برای توسعه دهندگان وب ایجاد می کند، مهم نیست که قبلا چند بار این کار را انجام داده باشید، چون در هر پروژه کاری شما، شرایط متفاوت است و نیاز به تحلیل و بررسی دقیق تر از اهداف خود دارید.

وقتی با برنامه نویسان در مورد وب اسکریپت صحبت می کنم، سوء تفاهم و سردرگمی زیادی در مورد این مورد وجود دارد. برخی از افراد مطمئن نیستند که قانونی است و یا نحوه برخورد با مشکلاتی همچون صفحاتی که با جاوا اسکریپت بارگزاری می شوند یا نوع رفتار اسکریپت با صفحات ورود به سایت ها چگونه است. بسیاری در مورد چگونگی شروع یک پروژه بزرگ وب اسکریپت یا حتی محل پیدا کردن داده هایی که به دنبال آن هستند سردرگم هستند. این کتاب به دنبال پایان دادن به بسیاری از این سؤالات رایج و تصورات غلط در مورد وب اسکریپت است که راهنمای جامعی برای اکثر پروژه های متداول وب اسکریپت ارائه می دهد. این حوزه یک زمینه متنوع و سریع در حال تغییر است، و ما سعی کرده ایم مفاهیم سطح بالا و مثال های عینی را برای پوشش تقریبا هر پروژه جمع آوری داده ارائه دهیم.

در سراسر کتاب، نمونه کدهایی برای مفهوم هرچه دقیق تر مطالب ارائه شده است و به شما امکان می دهد آن ها را پیاده سازی کنید.

نیما شفیعی رضوانی نژاد

افشین اسمعیل زاد