



به نام خدا

وب اسکرپینگ

با پایتون

مؤلفان:

نیما شفیعی رضوانی نژاد

بهاره بهروز



هرگونه 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 اسکرپینگ با پایتون

◀ مولفان: نیما شفیعی رضوانی نژاد - بهاره بهروز

◀ ناشر: موسسه فرهنگی هنری دیباگران تهران

◀ ویراستار: مهدیه مخبری

◀ صفحه آرای: نازنین نصیری

◀ طراح جلد: داریوش فرسایی

◀ نوبت چاپ: اول

◀ تاریخ نشر: ۱۴۰۲

◀ چاپ و صحافی: نامن

◀ تیراژ: ۱۰۰ جلد

◀ قیمت: ۲۰۵۰۰۰۰ ریال

◀ شابک: ۹۷۸-۶۲۲-۲۱۸-۷۹۱-۰

◀ 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 دیباگران تهران :

WWW.MFTBOOK.IR

www.dibagaran-tehran.com

سرشناسه: شفیعی رضوانی نژاد، نیما، ۱۳۷۱-
عنوان و نام پدیدآور: وب اسکرپینگ با پایتون /
مولفان: نیما شفیعی رضوانی نژاد، بهاره بهروز؛
ویراستار: مهدیه مخبری.
مشخصات نشر: تهران: دیباگران تهران: ۱۴۰۲
مشخصات ظاهری: ۱۹۰ ص: مصور،
شابک: ۹۷۸-۶۲۲-۲۱۸-۷۹۱-۰
وضعیت فهرست نویسی: فیپا
موضوع: پایتون (زبان برنامه نویسی کامپیوتر)
موضوع: Python (computer program language)
موضوع: استخراج داده های وب
موضوع: Web scraping
شناسه افزودن: بهروز، بهاره، ۱۳۶۶-
رده بندی کنگره: ۲/۷۶/۷۳ QA
رده بندی دیویی: ۰۰۵/۱۳۳
شماره کتابشناسی ملی: ۹۴۶۹۵۹۶

نشانی اینستاگرام دیبا dibagaran_publishing نشانی تلگرام: @mftbook

هر کتاب دیباگران، یک فرصت جدید علمی و شغلی.

هر گوشی همراه، یک فروشگاه کتاب دیباگران تهران.

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

فهرست مطالب

۷	مقدمه ناشر
۸	پیشگفتار
۸	وب اسکرپینگ چیست؟
۹	چرا وب اسکرپینگ؟
۱۱	درباره این کتاب

بخش اول

۱۲	ساختار خراش دهنده ها
----	----------------------------

۱۴	فصل ۱ وب اسکرپینگ چیست؟
۱۵	چه کسی از وب اسکرپینگ استفاده می کند؟
۱۶	اولین وب خراش شما
۱۷	معرفی ابزارهای خزش وب

بخش دوم

۲۸	اسکرپینگ پیشرفته وب
----	---------------------------

۳۰	فصل ۲ خواندن اسناد
۳۰	رمزگذاری فایل ها
۳۱	متن
۳۵	CSV
۳۷	PDF
۳۹	Microsoft word و .docx

۴۳	فصل ۳ فرایند تمیزسازی داده های آلوده
۴۶	تطبیق داده ها (Data Normalization)
۴۸	تمیز کردن داده ها
۵۰	پاکسازی داده ها

۵۲	فصل ۴ خواندن و نوشتن زبان های طبیعی
۵۲	خلاصه سازی داده ها
۵۶	مدل های مارکوف
۶۲	تجزیه و تحلیل آماری با استفاده از NLTK

۶۵ تحلیل واژگانی با استفاده از NLTK

۶۸ منابع یادگیری اضافی

فصل ۵ پیمایش از طریق فرم‌ها و ورود به سیستم ۶۹

۶۹ کتابخانه requests پایتون

۷۳ ارسال فایل‌ها و تصاویر

۷۴ مدیریت ورود و کوکی‌ها

۷۵ احراز هویت دسترسی اصلی HTTP

۷۶ مشکلات دیگر فرم

فصل ۶ اسکرپینگ جاوااسکریپت ۷۷

۷۷ جاوااسکریپت

۷۷ مقدمه کوتاهی به جاوااسکریپت

۷۹ کتابخانه‌های متداول جاوااسکریپت

۸۲ اجرای جاوااسکریپت در پایتون با Selenium

۸۵ مدیریت انتقال‌ها

۸۷ یک نکته آخر در مورد جاوااسکریپت

فصل ۷ گشت‌زنی از طریق APIs ۸۸

۸۸ معرفی مختصری از API‌ها

۹۰ متدهای HTTP و API‌ها

۹۱ اطلاعات بیشتر در مورد پاسخ‌های API

۹۲ تجزیه و تحلیل JSON

۹۴ API‌های غیر مستند

۹۷ یافتن و مستند کردن API‌ها به صورت خودکار

۱۰۰ ترکیب API‌ها با منابع داده دیگر

۱۰۴ بیشتر در مورد API‌ها

فصل ۸ پردازش تصویر و تشخیص متن ۱۰۵

۱۰۶ مروری بر کتابخانه‌ها

۱۰۹ پردازش متن با قالب‌بندی خوب

۱۱۲ تنظیم تصاویر به صورت خودکار

۱۱۵ استخراج متن از تصاویر در وبسایت‌ها

۱۱۸ آموزش Tesseract و خواندن CAPTCHA‌ها

۱۲۰ آموزش Tesseract

۱۲۳ دریافت CAPTCHA و ارسال راه‌حل‌ها

فصل ۹ جلوگیری از تله‌های اسکرپینگ (Scraping Traps) ۱۲۶

- ۱۲۶..... یک نکته در مورد اخلاق
- ۱۲۷..... شبیه انسان
- ۱۲۷..... تنظیم هدرهای HTTP
- ۱۲۹..... مدیریت کوکی‌ها با جاوا اسکریپت
- ۱۳۱..... زمان‌بندی مهم است
- ۱۳۲..... ویژگی‌های مشترک امنیتی فرم‌ها
- ۱۳۲..... مقادیر فیلد ورودی مخفی
- ۱۳۳..... جلوگیری از تله‌گره‌ها
- ۱۳۶..... چک‌لیست انسانی

فصل ۱۰ تست وب‌سایت خود با اسکرپرها ۱۳۸

- ۱۳۸..... تعریف Unit test
- ۱۳۹..... ماژول unittest پایتون
- ۱۴۱..... آزمایش ویکی‌پدیا
- ۱۴۵..... آزمایش با Selenium
- ۱۴۵..... تعامل با سایت
- ۱۴۸..... گرفتن عکس‌های صفحه
- ۱۴۹..... Selenium یا unittest؟

فصل ۱۱ پیمایش وب به صورت موازی ۱۵۱

- ۱۵۱..... فرآیندها در مقابل نخ‌ها
- ۱۵۲..... کراولینگ چندنخی
- ۱۵۵..... شرایط رقابتی و صف‌ها
- ۱۵۸..... Threading
- ۱۶۰..... ماژول پردازش
- ۱۶۲..... وب کراولینگ چند پردازشی
- ۱۶۴..... ارتباط بین فرآیندها
- ۱۶۶..... کراولینگ چندپردازشی - رویکرد دیگری

فصل ۱۲ جمع‌آوری اطلاعات از راه دور ۱۶۸

- ۱۶۸..... چرا از سرورهای از راه دور استفاده کنید؟
- ۱۶۸..... جلوگیری از مسدودسازی آدرس IP
- ۱۷۰..... قابلیت حمل و توسعه‌پذیری
- ۱۷۰..... Tor
- ۱۷۱..... (PySocks)
- ۱۷۲..... میزبانی از راه دور

فصل ۱۳ قوانین و اخلاق وب کاوی ۱۷۶

۱۷۶.....علائم تجاری، حق نشر، پتنت

۱۷۷.....قانون کپی رایت

۱۷۹.....نفوذ به حریم متعلق به دیگران

۱۸۱.....قانون تقلب و سوءاستفاده کامپیوتری

۱۸۱.....robots.txt و شرایط خدمات

۱۸۵.....سه اسکرپر وب

۱۹۰.....حرکت به جلو

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 دست دارید تألیف "جناب آقای نیما شفیعی رضوانی نژاد و سرکار خانم بهاره بهروز" 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

پیشگفتار

تابه حال برنامه‌های کاربردی زیادی بر روی گوشی و کامپیوتر خودتان دیده‌اید، اما آیا در مورد نحوه توسعه این برنامه‌ها و اهدافشان تحقیق کرده‌اید؟ هدف برنامه‌نویسی، راحت‌تر و سریع انجام شدن کار انسانهاست.

وب اسکرپینگ مفهومی است که به برنامه‌نویسی روح و زندگی می‌دهد؛ در صفحات مختلف می‌چرخد و اطلاعاتی که مورد نظر ما است را به ما برمی‌گرداند. به نظر بسیار هیجان‌انگیز است.

روح تازه‌ای که وب به کارهایمان می‌دهد قابل توصیف نیست.

متأسفانه بسیاری از برنامه‌نویسان، دچار سوءتفاهم و ابهاماتی در مورد عملکرد وب اسکرپینگ می‌شوند. برخی افراد مطمئن نیستند که این کار، یک کار قانونی است یا نه ممکن است آنها ندانند چگونه مشکلات صفحات سنگین جاوااسکریپت را مدیریت کنند یا فایل‌های مورد نیاز را مدیریت کنند. بسیاری از آنها در مورد چگونگی شروع یک پروژه بزرگ وب اسکرپینگ یا حتی منابع داده‌هایی که به دنبالشان هستند، دچار سردرگمی شده‌اند. این کتاب به دنبال خاتمه دادن به بسیاری از این سؤال‌های متداول و سوءتفاهم‌ها در مورد وب اسکرپینگ است و اگر تا انتهای کتاب آن را بخوانید پاسخ بسیاری از سوالاتتان رفع می‌شود.

وب اسکرپینگ یک حوزه متفاوت و جدید است و ما تلاش کرده‌ایم تا مفاهیم پیشرفته و نمونه‌های واقعی از پروژه‌ها را ارائه دهیم تا بتوانید هر پروژه جمع‌آوری داده‌ای که احتمالاً با آن روبه‌رو خواهید شد، را انجام دهید. در طول کتاب، نمونه‌هایی برای نمایش این مفاهیم ارائه شده‌اند و به شما اجازه داده می‌شود تا آنها را امتحان کنید.

وب اسکرپینگ چیست؟

اگرچه وب اسکرپینگ تکنولوژی جدیدی نیست و سابقه آن تقریباً با ظهور اینترنت برابری می‌کند، اما در سال‌های گذشته این کار بیشتر با کاربردهای وب خراش^۱، استخراج داده‌ها^۲، برداشتن اطلاعات از وب^۳ یا موارد مشابه شناخته می‌شد.

ما در این کتاب بر مفاهیم وب اسکرپینگ و خزنده‌های وب^۴ اشاره می‌کنیم و به دفعات از این عبارات در طول کتاب استفاده می‌کنیم.

اگر بخواهیم تعریفی از وب اسکرپینگ داشته باشیم اینطور می‌توانیم بگوییم؛ وب اسکرپینگ کاری است که انسانها از طریق آن از یک مرورگر وب داده‌هایی را جمع‌آوری می‌کنند. این کار توسط برنامه‌ای که از یک وب سرور استعلام می‌گیرد و داده‌ها را درخواست می‌کند، به صورت خودکار انجام می‌شود، خروجی بدست آمده معمولاً

1- Screen Scraping
2- Data Extraction
3- web harvesting
4- web crawlers

به صورت HTML و سایر فایل‌هایی که صفحات وب را تشکیل می‌دهند، می‌باشد و در مرحله بعدی آن داده‌ها را جهت استخراج اطلاعات مورد نیاز تجزیه و تحلیل می‌کند.

در حالت کلی، وب اسکرپینگ شامل مجموعه‌ای گسترده از تکنیک‌ها و فناوری‌های برنامه‌نویسی، مانند تجزیه و تحلیل داده، تجزیه و تحلیل زبان طبیعی و امنیت اطلاعات است.

به دلیل گسترده بودن دامنه این علم، در این کتاب مبانی وب اسکرپینگ و وب کراولینگ را در قسمت اول و مباحث پیشرفته را در قسمت دوم بررسی می‌کنیم. پیشنهاد ما این است که همه خوانندگان با دقت از مباحث مقدماتی تا پیشرفته کتاب را بررسی کنند تا مباحث را به طور کامل متوجه شوند.

چرا وب اسکرپینگ؟

مرورگرها برای اجرای کدهای جاوااسکریپت و نمایش تصاویر و ترتیب اشیاء در یک قالب قابل درک برای انسان مناسبند، اما آنها به تنهایی قادر نیستند طیف گسترده‌ای از امکانات وب را در اختیار شما قرار دهند.

وب اسکرپرها در جمع‌آوری و پردازش مقدار بزرگی از داده، سرعت و دقت بالایی دارند. آنها به جای باز کردن تک به تک صفحات، در یک چشم به هم زدن می‌توانند پایگاه‌های داده‌ای با هزاران یا حتی میلیون‌ها صفحه را به طور همزمان مشاهده و نتایج آن را در اختیار شما قرار دهند.

علاوه بر این، وب اسکرپرها به جاهایی می‌توانند دسترسی پیدا کنند که موتورهای جستجوی سنتی نمی‌توانند به راحتی به آنها دسترسی یابند. اگر کلمه‌ای مانند "بهترین پروازها به کیش" را در گوگل جستجو کنید، تعدادی از تبلیغات و سایت‌های محبوب پرواز برای شما نشان داده می‌شود. درحالی‌که گوگل فقط درباره محتوای این وبسایت‌ها می‌داند، نه نتایج دقیق جستجوی‌های مختلفی که در یک برنامه جستجو وارد شده است. با این حال، یک وب اسکرپر کاملاً توسعه یافته می‌تواند هزینه پرواز به کیش را در یک بازه زمانی، از بین میلیون‌ها وبسایت، به شما نمایش دهد و به شما بهترین زمان خرید بلیط را بگوید.

سؤالی که پیش می‌آید: "آیا جمع‌آوری داده‌ها مخصوص API ها است؟" اگر یک API مناسب را بیابید که با اهداف شما سازگار باشد متوجه قدرت آنها خواهید بود. آنها به صورتی طراحی شده‌اند تا جریان مناسبی از داده را به فرمت منظمی از یک برنامه کامپیوتری به برنامه کامپیوتر دیگر، ارائه دهند. از یک API می‌توانید برای بسیاری از انواع داده‌ها استفاده کنید، مانند پست‌های توییتر یا صفحات ویکی‌پدیا. به طور کلی، داده‌های بدست آمده از یک API اطلاعات بهتری نسبت به ساخت یک ربات برای گرفتن همان داده را به ما می‌دهد. با این حال، ممکن است هیچ API وجود نداشته باشد یا در جهت رفع مشکل شما نباشد، به عنوان مثال زمانیکه بخواهید از یک یا چند وبسایتی که هیچ API ای ندارند داده‌هایی را جمع‌آوری کنید، وب اسکرپرها می‌توانند ابزار قدرتمندی باشند.

به عنوان مثال دیگر، فرض کنید که شما می‌خواهید داده‌های جغرافیایی مانند اطلاعات فروشگاه‌ها یا مکان‌های جذاب دیگر را جمع‌آوری کنید. برخی از این داده‌ها از طریق API ها قابل دسترسی‌اند، اما بعضی از آنها فقط در صفحات وب در دسترس هستند. در اینجا، وب اسکرپرها می‌توانند به شما کمک کنند تا داده‌های مورد نظر را از منابعی که API ندارند، جمع‌آوری کنید.

همچنین، برنامه‌های وب اسکرپر می‌توانند به شما در جمع‌آوری داده‌هایی کمک کنند که در ابتدا حتی نمی‌دانید که به آنها نیاز دارید یا نه. یک برنامه وب اسکرپر می‌تواند به شما کمک کند تا درک بهتری از داده‌های موجود در وب پیدا کنید و ایده‌های جدیدی در ذهنتان پرورش دهید. به‌عنوان مثال، یک وب اسکرپر در شناسایی ترندهای موجود در شبکه‌های اجتماعی، عناوین روزنامه‌ها یا تغییرات در نظرسنجی‌ها به شما کمک می‌کند.

درنهایت، وب اسکرپینگ یک مهارت برنامه‌نویسی قدرتمند است که می‌تواند برای اهداف مختلفی مورد استفاده قرار گیرد. اگرچه این کتاب به شما اصول اساسی و تکنیک‌های وب اسکرپینگ را آموزش می‌دهد، اما درخواستی که از شما داریم این است که از این دانش برای اهداف قانونی و اخلاقی استفاده کنید. در فصل ۱ در مورد قوانین مربوط به وب اسکرپینگ، حقوق مالکیت و حفاظت از داده‌ها صحبت می‌کنیم. انتظاری که می‌رود این است که این مسائل را به دقت بررسی کنید.

درباره این کتاب

این کتاب نه تنها به عنوان مقدمه‌ای برای خراش دادن وب^۱ طراحی شده است، بلکه به عنوان یک راهنمای جامع برای جمع‌آوری، تبدیل و استفاده از داده‌ها از منابع دیگری که از زبان برنامه‌نویسی پایتون استفاده می‌کند، در نظر گرفته می‌شود. بنابراین نمی‌توان آن را به عنوان کتابی با اطلاعات پایه فقط در نظر گرفت در این کتاب بسیاری از اصول اولیه پایتون پوشش داده می‌شود.

پس حتی اگر پایتون کار نکرده باشید هم این کتاب برای شما بسیار مناسب است. ما سعی کرده‌ایم تمام مفاهیم و نمونه‌های کد را در سطح برنامه‌نویسی پایتون از ابتدا تا سطح متوسط ارائه دهیم.

این مفاهیم برای طیف وسیعی از خوانندگان قابل فهم است.

خراش دادن وب یک موضوع گسترده است شامل، استفاده از پایگاه‌های داده^۲، وب سرورها^۳، HTML، HTTP، امنیت اینترنت^۴، پردازش تصویر^۵، علم داده^۶ و ابزارهای دیگر که در این کتاب سعی داریم همه موارد را پوشش دهیم.

در بخش اول موضوع خراش دادن وب و خزیدن وب^۷ به صورت کامل پوشش داده می‌شود. در این بخش تمرکز ما بر تعداد زیادی از کتابخانه‌های مورد استفاده در سراسر کتاب است. قسمت اول را به جرأت می‌توانیم به عنوان یک مرجع جامع برای این کتابخانه‌ها و تکنیک‌ها بنامیم. مهارت‌های آموزش داده شده در بخش اول احتمالاً برای همه مفید خواهد بود.

در بخش دوم به نوشتن اسکریپت‌های وب می‌پردازیم.

ساختار این کتاب طوری طراحی شده تا بتوانید سریعاً جواب سؤال خود را از صفحات مورد نظر پیدا کنید.

-
- 1- Web scraping
 - 2- Data base
 - 3- Web servers
 - 4- internet security
 - 5- Image processing
 - 6- data science
 - 7- web crawling