به نام خدا

وب اسکرپینگ

باپایتون

مؤلفان:

نیما شفیعی رضوانی نژاد

بهاره بهروز

مؤسسه فرهنگی هنری
دیباگران تهران

عنوان کتاب: وب اسکرپینگ با پایتون

مولفان : نیما شفیعی رضوانی نژاد-بهاره بهروز

ناشر: موسسه فرهنگی هنری دیباگران تهران
ویراستار: مهدیه مخبری
صفحه آرایی: نازنین نصیری
طراح جلد:داریوش فرسایی
نوبت چاپ: اول
تاریخ نشر: ۱۴۰۲
چاپ و صحافی:ثامن
تیراژ:۱۰۰ جلد
قیمت:۲۰۵۰۰۰۰ ریال
شابک: ۰-۷۹۱-۲۱۸-۶۲۲-۹۷۸

نشانی واحد فروش:تهران، خیابان انقلاب،خیابان دانشگاه -تقاطع شهدای ژاندارمری-پلاک ۱۵۸ساختمان دانشگاه- طبقه دوم-واحد۴ تلفن ها: ۶۶۹۶۵۷۴۹-۲۲۰۸۵۱۱۱

فروشگاههای اینترنتی دیباگران تهران :
WWW.MFTBOOK.IR
www.dibagarantehran.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: ۰-۷۹۱-۲۱۸-۶۲۲-۹۷۸
وضعیت فهرست نویسی: فیپا
موضوع:پایتون(زبان برنامه نویسی کامپیوتر)
موضوع:Python(computer program language)
موضوع:استخراج داده های وب
موضوع:Web scraping
شناسه افزوده:بهروز،بهاره،۱۳۶۶-
رده بندی کنگره:۲پ/۷۶/۷۳ QA
رده بندی دیویی:۰۰۵/۱۳۳
شماره کتابشناسی ملی:۹۴۶۹۵۹۶

نشانی تلگرام:@mftbook نشانی اینستاگرام دیبا dibagaran_publishing

فهرست مطالب

مقدمه ناشر

خط مشی انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب هایی با کیفیت عالی است که بتواند خواسته های 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.

گستردگی علوم وسرعت 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 زمینه های گوناگون و مورد نیاز جامعه تلاش نموده برای رفع کمبودها و نیازهای موجود، منابعی پُربار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهد.

کتابی که در دست دارید تألیف **"جناب آقای نیما شفیعی رضوانی نژادو سرکار خانم بهاره بهروز"** است که با تلاش همکاران ما درنشر دیباگران تهران منتشرگشته و شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

با نظرات خود مشوق و راهنمای ما باشید

با ارائه نظرات و پیشنهادات وخواسته های خود،به ما کمک کنید تا بهتر و دقیق تر در جهت رفع نیازهای علمی و آموزشی کشورمان قدم برداریم.برای رساندن پیام هایتان به ما از رسانه های دیباگران تهران شامل سایتهای فروشگاهی و صفحه اینستاگرام و شماره های تماس که در صفحه شناسنامه کتاب آمده استفاده نمایید.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
dibagaran@mftplus.com

پیشگفتار

تابه‌حال برنامه‌های کاربردی زیادی بر روی گوشی و کامپیوتر خودتان دیده‌اید، اما آیا در مورد نحوهٔ 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 پروژهٔ جمع‌آوری داده‌ای که احتمالاً با آن روبه‌رو خواهید شد، را انجام دهید. در طول کتاب، نمونه کدهایی 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[1]، استخراج داده‌ها[2]، برداشتن اطلاعات از وب[3] یا موارد مشابه شناخته می‌شد.

ما در این کتاب بر مفاهیم وب اسکرپینگ و خزنده‌های وب[4] 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، به‌صورت خودکار انجام می‌شود، خروجی بدست آمده معمولاً

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

به‌صورت HTML و سایر فایل‌هایی که صفحات وب را تشکیل می‌دهند، می‌باشد و در مرحله بعـدی آن داده‌هـا را جهت استخراج اطلاعات مورد نیاز تجزیه‌وتحلیل می‌کند.

در حالت کلی، وب اسکرپینگ شامل مجموعه‌ای گسترده از تکنیک‌هـا و فنـاوری‌هـای برنامـه‌نویسـی، ماننـد تجزیه‌وتحلیل داده، تجزیه‌وتحلیل زبان طبیعی و امنیت اطلاعات است.

به دلیل گسترده بودن دامنه این علم، در این کتاب مبانی وب اسکرپینگ و وب کراولینگ را در قسـمت اول و مباحث پیشرفته را در قسمت دوم بررسی می‌کنیم. پیشنهاد ما این است که همـه خواننـدگان بـا دقـت از مباحـث مقدماتی تا پیشرفتهٔ کتاب را بررسی کنند تا مباحث را به‌طور کامل متوجه شوند.

چرا وب اسکرپینگ؟

مرورگرها برای اجرای کدهای جاوااسکریپت و نمایش تصاویر و ترتیب اشیاء در یک قالب قابل درک برای انسـان 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 بـازکردن تک به تک صفحات، در یک چشم به هم زدن می‌توانند پایگاه‌های داده‌ای با هزاران یا حتی میلیون‌ها صـفحه را به‌طور همزمان مشاهده و نتایج آن را در اختیار شما قرار دهند.

علاوه‌بر این، وب اسکرپرها به جاهایی می‌توانند دسترسی پیدا کنند که موتورهای جستجوی سنتی نمـی‌تواننـد به راحتی به آنها دسترسی یابند. اگر کلمه‌ای مانند"بهترین پروازها به کیش"را در گوگل جستجو کنید، تعـدادی از تبلیغات و سایت‌های محبوب پرواز برای شما نشان داده می‌شود. درحالی‌کـه گوگـل فقـط دربـاره محتـوای ایـن وبسایت‌ها می‌داند، نه نتایج دقیق جستجوی‌های مختلفی که در یک برنامه جستجو وارد شده اسـت. بـااین‌حـال، یک وب اسکرپر کاملاً توسعه یافته می‌تواند هزینه پرواز به کیش را در یک بازه زمانی، از بین میلیونها وب‌سـایت، 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 بیابیـد کـه بـا اهداف شما سازگار باشد متوجه قدرت آنها خواهید بود. آنها به صورتی طراحی شده‌اند تا جریان مناسـبی 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. به‌طورکلی، داده‌های بدست آمده از یک API اطلاعات بهتری نسبت به ساخت یک ربات برای گرفتن همان داده را به ما می‌دهد. بااین‌حال، ممکن اسـت هیچ API وجود نداشته باشد یا در جهت رفع مشکل شما نباشد، به‌عنوان مثال زمانیکه بخواهیـد از یـک یـا چنـد وبسایتی که هیچ API ای ندارند داده‌هایی را جمع‌آوری کنید، وب اسکرپرها می‌توانند ابزار قدرتمندی باشند.

به‌عنوان 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 آنهـا فقـط در صفحات وب در دسترس هستند. در اینجا، وب اسکرپرها می‌توانند به شما کمک کنند تا داده‌های مورد نظـر 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 پروش 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 به‌عنوان مقدمه‌ای برای خراش دادن وب[1] طراحی شده است، بلکه به‌عنوان یک راهنمای جامع برای جمع‌آوری، تبدیل و استفاده از داده‌ها از منابع دیگری که از زبان برنامه‌نویسی پایتون استفاده می‌کند، درنظر گرفته می‌شود. بنابراین نمی‌توان آن را به‌عنوان کتابی با اطلاعات پایه فقط درنظر 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[2]، وب سرورها[3]، HTTP، HTML، امنیت اینترنت[4]، پردازش تصویر[5]، علم داده[6] و ابزارهای دیگرکه در این کتاب سعی داریم همه موارد را پوشش دهیم.

در بخش اول موضوع خراش دادن وب و خزیدن وب[7] به‌صورت 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 به‌عنوان 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 اسکریپرهای وب می‌پردازیم.

ساختار این کتاب طوری طراحی شده تا بتواتید 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