
*به نام خدا*

# آموزش وب کاوی با

# SCRAPY

**مؤلف:**

**محمدرضا شاقوزی**



◂ عنوان کتاب: **آموزش وب کاوی با SCRAPY**

◂ مولف : **محمدرضا شاقوزی**

◂ ناشر: **موسسه فرهنگی هنری دیباگران تهران**

◂ **ویراستار:** مهدیه مخبری

◂ **صفحه آرایی:** نازنین نصیری

◂ **طراح جلد:** داریوش فرسایی

◂ **نوبت چاپ:** اول

◂ **تاریخ نشر:** ۱۴۰۱

◂ **چاپ و صحافی:** درج عقیق

◂ **تیراژ:** ۱۰۰ جلد

◂ **قیمت:** ۱۰۲۰۰۰۰ ریال

◂ **شابک:** ۹۷۸-۶۲۲-۲۱۸-۶۴۸-۷

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

**فروشگاههای اینترنتی دیباگران تهران :**

**WWW.MFTBOOK.IR**

**www.dibagarantehran.com**

سرشناسه:شاقوزی،محمدرضا،۱۳۷۳-

عنوان و نام پدیدآور:آموزش وب کاوی با SCRAPY /مولف: محمدرضا شاقوزی؛

ویراستار:مهدیه مخبری.

مشخصات نشر: تهران : دیباگران تهران :۱۴۰۱

مشخصات ظاهری:۱۲۲ ص:جدول

شابک: ۷-۶۴۸-۲۱۸-۶۲۲-۹۷۸

وضعیت فهرست نویسی: فیپا

موضوع:داده کاوی Data Mining

موضوع:اسکرپی scrapy(software framework)

موضوع:داده های کلان big data

رده بندی کنگره:۷۶/۹ QA

رده بندی دیویی:۰۰۶/۳۱۲

شماره کتابشناسی ملی:۹۱۲۴۱۱۴

**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 محمدرضا شاقوزی"** است که با تلاش همکاران ما در نشر 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

**مقدمه**

در قرن حاضر با ظهور تکنولوژی‌های ارتباطات، شاهد انفجار اطلاعات هستیم. کمتر موضوعی پیدا می‌شود که برای آن محتواهای بسیاری بر روی اینترنت نتوان پیدا کرد. از طرفی می‌دانیم هرچه اطلاعات بیشتری داشته باشیم در تصمیم‌گیری‌های روزمره سود بیشتری نصیبمان می‌شود. فرض کنید می‌خواهید یک ماشین کارکرده بخرید. در حالت معمولی، شما از طریق دوستان و آشنایانتان از قیمت خودرو و خریدار مطلع می‌شوید. اما اگر بتوانید به صورت هدفمند از یک سایت خرید و فروش دست دوم، قیمت خودروی مورد نظر را در طول روزهای مختلف جمع‌آوری کنید، چه مقدار سود خواهید کرد؟ حتی شاید بتوانید با یک دقت مناسبی، قیمت فردای آن خودرو را نیز پیش‌بینی کنید. یا شاید به عنوان یک مشاور معتمد برای دوستان و آشنایانتان عمل کنید و از ضرر احتمالی آن‌ها جلوگیری کنید. **وب‌کاوی با اسکرپی** شما را برای ساخت چنین ابزارهایی راهنمایی می‌کند.

این کتاب از مباحث پایه شروع می‌کند و به پروژه‌های تقریبا کاربردی ختم می‌شود. برای راحتی ارتباط خواننده، بین فصل‌هایی که به مفاهیم نظری می‌پردازند، یک فصل کاربردی قرار داده شده است تا روند مطالعه کتاب فرسایشی نشود.

لازم به ذکر است که کدهای پروژه و مراحل انجام آن در مخزن گیت‌هاب مربوط به این کتاب در لینک زیر قابل دسترس است. پس اگر کیفیت چاپ کدها یا لینک‌هایی که به صورت precent encoding در متن نمایش داده شده‌اند(af٪٨c%d٨٪)، حس خوبی به شما نداد، چیزی را از دست نداده‌اید. زیرا می‌توانید تمام کدها را از طریق لینک زیر به تفکیک فصل مشاهده کنید:

https://github.com/softrebel/web_mining_with_scrapy

البته برای حداکثر یادگیری پیشنهاد می‌شود که به هیچ عنوان کدها را کپی نکنید و خودتان کدنویسی کنید. همچنین به علت کمبود فضا، بعضی از خطوط کد پایتون در کتاب با کاراکتر "\" به دو خط تقسیم شده و همچنان قابل اجرا است. این در حالی است  که در مخزن سعی شده است کدها به کاربرد نزدیک‌تر باشد. لذا این تفاوت را به منزله اشتباه بودن کدهای کتاب در نظر نگیرید.

به دلیل اینکه وب‌سایت‌های فعلی ایستا نیستند، لازم شد بعضی از تمرین‌ها و مثال‌ها را بر روی یک سایت مختص این کتاب پیاده‌سازی کنم که در طول زمان تغییری ایجاد نشود. دامنه زیر برای این کار تهیه شده است:

http://www.pycrawling.ir/

البته این ایراد برای پروژه‌ها نیز وارد است و ممکن است در سال آینده ساختار آن‌ها تغییر کند. در صورت تغییرات، در ویرایش‌های بعدی و در مخزن گیت، آن‌ها را لحاظ می‌کنیم. همچنین شما می‌توانید انتقادات و پیشنهادات خود را با رایانامه sh.mohammad۶۶@gmail.com با بنده درمیان بگذارید تا در ویرایش‌های بعدی در نظر گرفته شود.

در انتها لازم است از صبر و زحمات تمام دست‌اندرکاران مجموعه دیباگران تشکر نمایم.


ارادتمند

محمدرضا شاقوزی