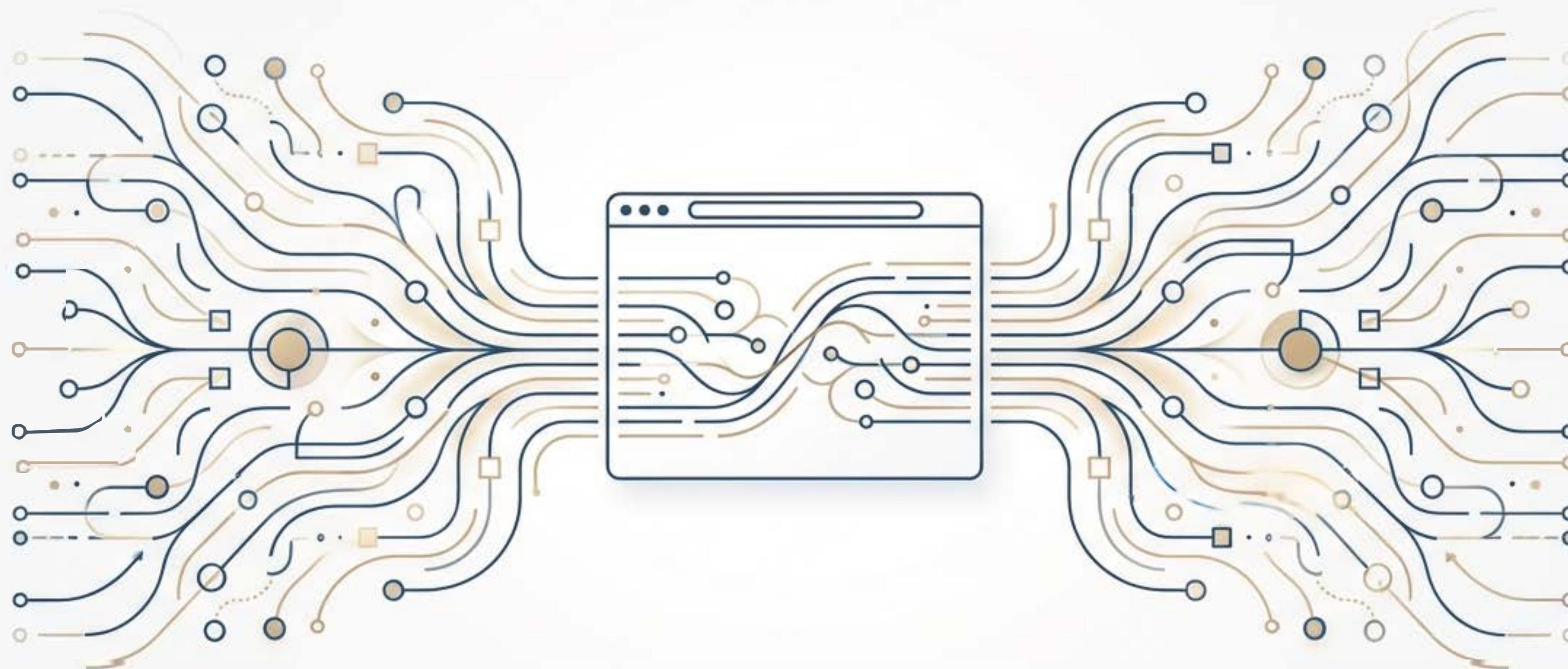
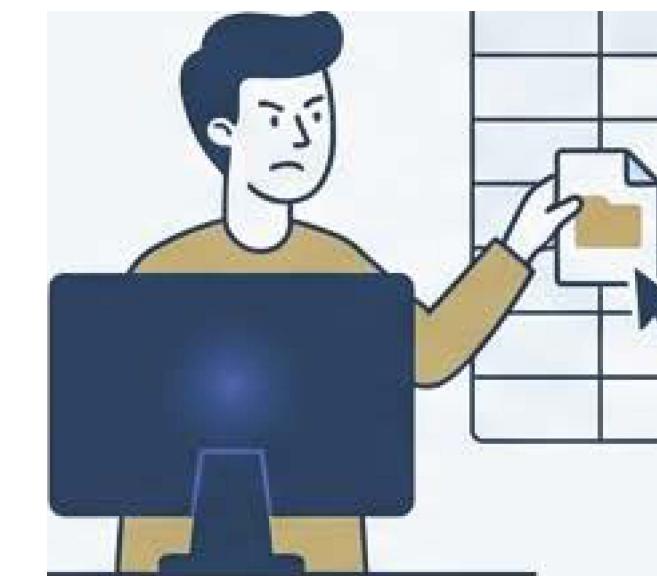


عنوان اصلی: وب، پایگاه داده‌ی شخصی شما: مقدمه‌ای بر هنر وب اسکرپینگ
زیرنویس: چگونه داده‌های وب را به مزیت رقابتی خود تبدیل کنیم.



عنوان: مشکل: اقیانوسی از داده، در بطری‌های شیشه‌ای

وب بزرگترین منبع داده در تاریخ بشر است:
قیمت محصولات، نظرات کاربران، اخبار، اطلاعات
تماس، روندهای بازار و...



اما این داده‌ها در ساختار بصری وبسایت‌ها "محبوس" شده‌اند.

استخراج دستی این اطلاعات کاری کند، خسته
کننده، و به شدت مستعد خطا است.

عنوان: راه حل: معرفی دستیار تحقیقاتی رباتیک شما

وب اسکرپینگ (Web Scraping) چیست؟ فرآیند استخراج خودکار و ساختاریافته‌ی داده‌ها از وبسایت‌ها.

به زبان ساده: یک ربات نرم‌افزاری که به جای شما وبسایت‌ها را "می‌خواند"، اطلاعات مورد نظر شما را پیدا می‌کند، و می‌کند، و آن‌ها را در یک فرمت قابل استفاده (مانند اکسل یا پایگاه داده) ذخیره می‌کند.



عنوان: از ایده تا واقعیت: قدرت وب اسکرپینگ در عمل

وب اسکرپینگ نیروی محرکه بسیاری از کسبوکارها و محصولات داده محور است.



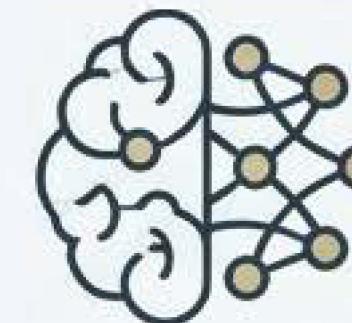
ساخت موتورهای مقایسه
قیمت و رصد بازار.



تحلیل انبوه نظرات کاربران برای
درک روندهای بازار.

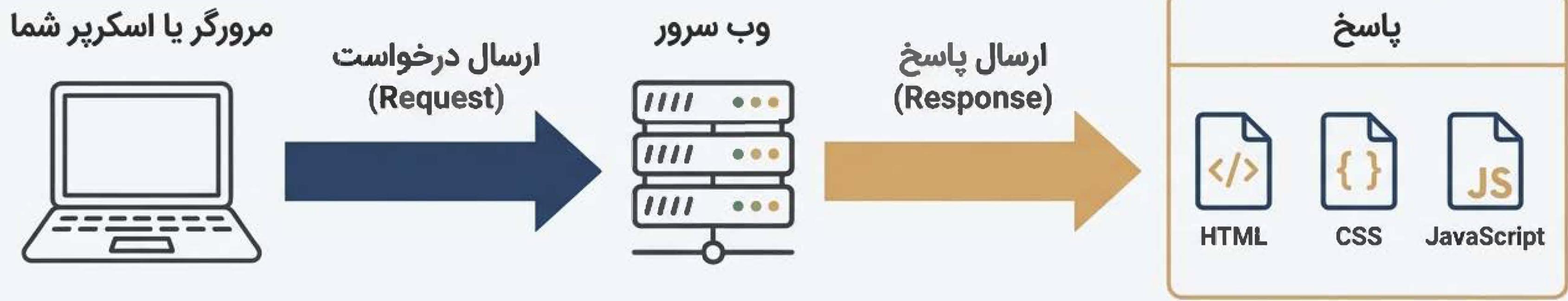


جمعآوری اطلاعات تماس
کسبوکارها برای بازاریابی و
فروش.



ساخت مجموعه داده‌های
اختصاصی برای آموزش مدل‌های
یادگیری ماشین.

عنوان: معماری وب در ۶۰ ثانیه: پیش‌نیاز اسکرپینگ



اسکرپر ما دقیقاً مانند یک مرورگر عمل می‌کند، اما به جای نمایش صفحه برای انسان، کد آن را برای تحلیل می‌خواند.

عنوان: نقشه گنج: پیدا کردن داده در کد HTML

```
<div class="product-card">
  
  <div class="product-details">
    <h1 style="color: #003366;">گوشی هوشمند مدل X</h1>
    <p class="description">آخرین مدل با بهترین امکانات...</p>
    <div class="price" style="color: #D4A26A;">
      ۱۹,۹۹۹,۰۰۰ تومان
    </div>
    <button>افزودن به سبد خرید</button>
  </div>
</div>
```



گوشی هوشمند مدل X
آخرین مدل با بهترین امکانات...
بهترین امکانات...

۱۹,۹۹۹,۰۰۰ تومان

افزودن به سبد خرید

مفهوم کلیدی: هر وبسایت یک ساختار قابل پیش‌بینی دارد. ما با استفاده از "آدرس‌ها" (Selectors)، به اسکرپر می‌گوییم دقیقاً کدام داده را بردارد.

عنوان: جعبه ابزار یک اسکرپر حرفه‌ای

انتخاب ابزار مناسب برای هر کار، کلید موفقیت است. در اکوسیستم پایتون، ابزارهای قدرتمندی برای این کار وجود دارد.



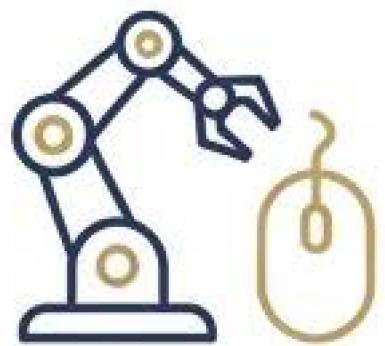
Requests

`Requests`: پیک شما. برای ارسال درخواست HTTP و دریافت کد خام صفحه.



Beautiful Soup

`Beautiful Soup`: نقشه‌خوان شما. برای تجزیه (Parse) کد HTML و جستجوی جستجوی آسان در آن.



Selenium

`Selenium`: راننده رباتیک شما. برای کنترل یک مرورگر واقعی و اسکرپ کردن سایتها پویا (JavaScript-heavy).



Scrapy

`Scrapy`: خط تولید شمل. یک فریمورک کامل برای ساخت اسکرپرهای بزرگ، سریع و پیچیده.

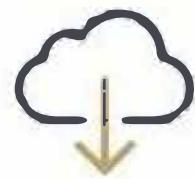
عنوان: دستور پخت یک اسکرپر ساده: فرآیند گام به گام



1.

(Request)

آدرس URL را به کتابخانه `requests` می‌دهیم.



2.

(Response)

کد کامل HTML صفحه را دریافت می‌کنیم.



3.

(Parse)

کد HTML را به `Beautiful Soup` می‌دهیم تا آن را به یک ساختار قابل جستجو تبدیل کند.



4.

(Extract)

با CSS Selectors، عناصر مورد نظر (قیمت، عنوان) را پیدا می‌کنیم.



5.

(Store)

داده‌های تمیز را در یک فایل CSV، اکسل یا پایگاه داده ذخیره می‌کنیم.

عنوان: کد در عمل: استخراج عنوان یک مقاله خبری

```
# Import required libraries
import requests
from bs4 import BeautifulSoup

# Step 1 & 2: Send request and get the page HTML
url = 'https://example-news-website.com/article'
response = requests.get(url)

# Step 3: Parse the HTML with Beautiful Soup
soup = BeautifulSoup(response.content, 'html.parser')

# Step 4: Find the main title (usually in an <h1> tag)
title_element = soup.find('h1')
title_text = title_element.text

# Step 5: Print the result
print(f"The title is: {title_text}")
```

عنوان: چالش‌های دنیای واقعی و راه حل‌های هوشمندانه

وبسایت‌ها همیشه ساده نیستند. در اینجا با چند چالش رایج و راه حل‌های آن‌ها آشنا می‌شویم:



محتوای پویا (JavaScript)
داده‌هایی که پس از بارگذاری کامل صفحه با جاوااسکریپت نمایش داده می‌شوند.

راه حل: استفاده از ابزارهایی مانند Selenium که یک مرورگر واقعی واقعی را کنترل می‌کنند.



صفحه‌بندی (Pagination)
داده‌ها در چندین صفحه پخش شده‌اند (مثلًاً صفحه ۱، ۲، ۳...).

راه حل: طراحی یک حلقه (loop) در کد برای پیمایش خودکار تمام صفحات.



سدود شدن (Blocking)
وبسایت‌ها اسکریر شما را سدود می‌کنند.

راه حل: استفاده از هدرهای مناسب (User-Agents)، پراکسی‌ها، و رعایت فاصله زمانی منطقی بین درخواست‌ها.

عنوان: اسکرپر اخلاقمدار: قوانین طلایی یک شهروند خوب وب

قدرت زیاد، مسئولیت زیادی به همراه دارد. همیشه با احترام با وبسایت‌ها رفتار کنید.



به سرور فشار نیاورید

این فایل، قوانین دسترسی ربات‌ها به سایت را مشخص می‌کند. همیشه آن را بررسی کنید.



به سرور فشار نیاورید

بین درخواست‌های خود فاصله زمانی قرار دهید تا باعث کندی سایت نشود.



خودتان را معرفی کنید

در هدر درخواست (User-Agent)، هویت اسکرپر خود را مشخص کنید (مثلًا ("MyAwesomeProject Scraper")).

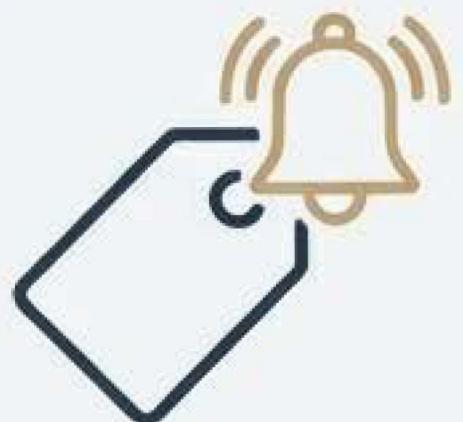


از داده‌ها مسئولانه استفاده کنید

به حق کپیرایت، حریم خصوصی کاربران کارگران و شرایط استفاده از وبسایت احترام بگذارید.

عنوان: پروژه بعدی شما چه خواهد بود؟

این مهارت جدید را برای حل مشکلات و ساخت ابزارهای شخصی به کار بگیرید.



ردیاب قیمت

اسکرپری بسازید که قیمت محصول مورد علاقه‌تان را کرده و در صورت تخفیف به شما اطلاع دهد.



داشبورد خبری شخصی

اخبار و مقالات را از منابع مورد علاقه‌تان جمع‌آوری کرده و در یک صفحه نمایش دهید.



تحلیل بازار محلی

قیمت مسکن یا خودرو را در منطقه خودتان از سایتها آگهی استخراج و تحلیل کنید.

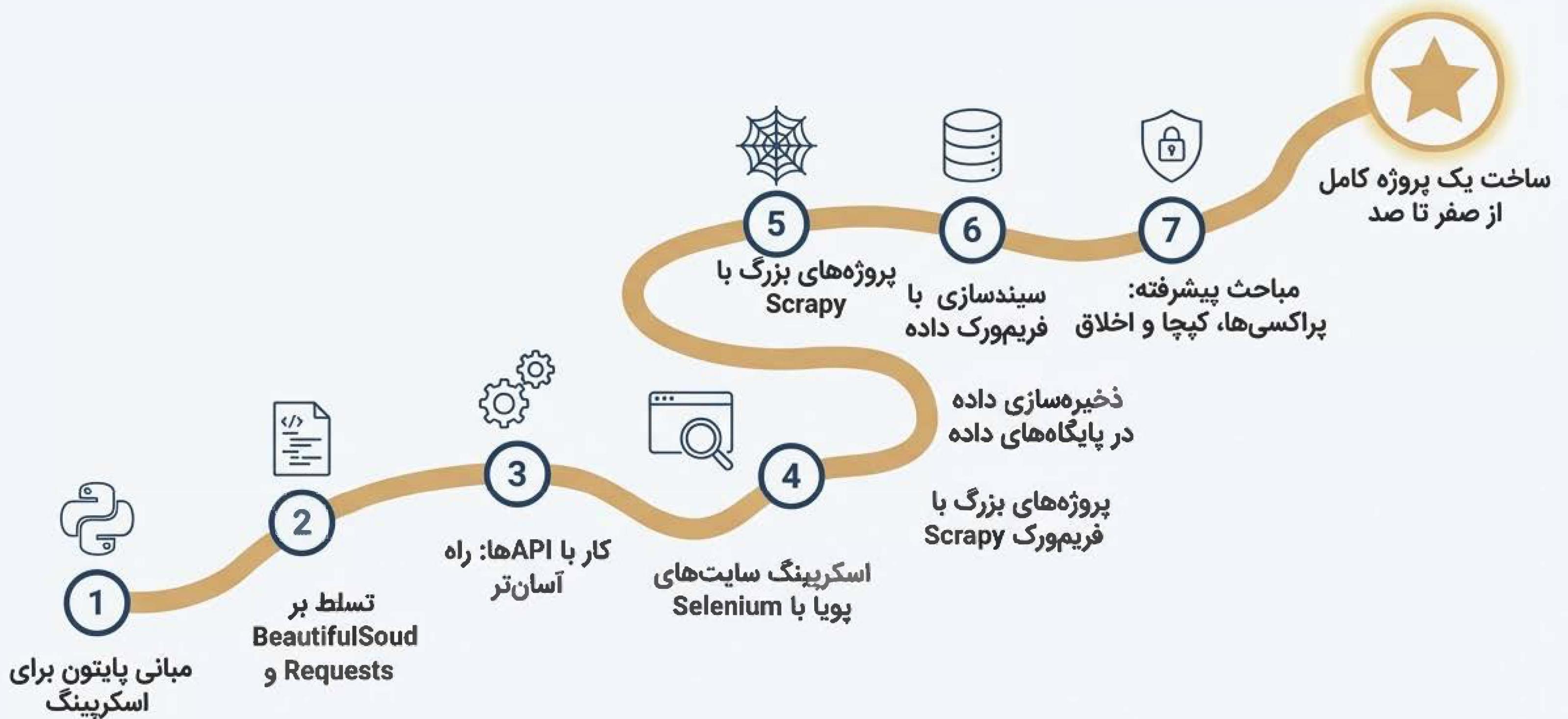


ساخت پورتفولیو

یک پروژه داده محور چشمگیر چشمگیر بسازید و آن را در رزومه و پروفایل لینکدین خود به نمایش بگذارید.

عنوان: نقشه راه تسلط: مسیر شما در دوره جامع وب اسکرپینگ

این جلسه تنها یک نقطه شروع بود. در دوره جامع، ما عمیقاً به هر یک از این مفاهیم و فراتر از آن خواهیم پرداخت.



عنوان: سوالات متدائل

چند سوال که ممکن است برایتان پیش آمده باشد:

سوال ۱: آیا وب اسکرپینگ قانونی است؟

پاسخ: بله، اگر به صورت اخلاقی و مسئولانه انجام شود. استخراج داده‌های عمومی معمولاً قانونی است، اما همیشه باید به فایل `robots.txt` و شرایط استفاده از وبسایت احترام گذاشت. ما در دوره به تفصیل این موضوع را بررسی می‌کنیم.

سوال ۲: برای شروع به چه سطح دانش پایتون نیاز دارم؟

پاسخ: آشنایی با مفاهیم پایه پایتون (متغیرها، حلقه‌ها، توابع) کافی است. مأذول اول دوره برای مرور و تقویت همین مبانی طراحی شده است.

سوال ۳: سخت‌ترین بخش وب اسکرپینگ چیست؟

پاسخ: معمولاً، مقابله با وبسایت‌هایی که به‌طور فعال با اسکرپینگ مبارزه می‌کنند (با استفاده از CAPTCHA یا تغییر مدام ساختار HTML) در دوره، استراتژی‌های مقابله با این چالش‌ها را یاد خواهید گرفت.

