



Web Scraping

Von Monica De Blasiis

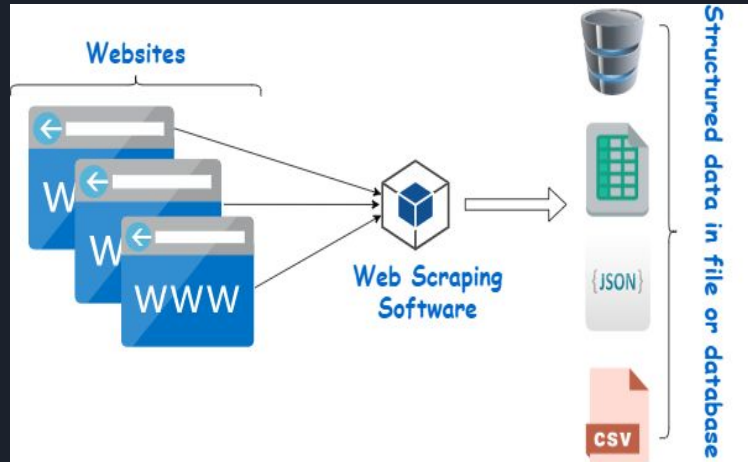


Inhaltsverzeichnis

- Was ist Web Scraping?
- Mein Projekt
- Warum wird Web Scraping blockiert?
- Wie kann Scraping eine Website beschädigen?
- Wie blockieren Webseiten Scraper?

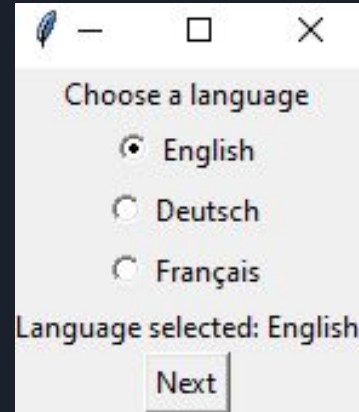
Was ist Web Scraping?

- Datenextraktion aus Webseiten
- Herunterladen und Parsen von HTML-Code
- Verwendungen: Website-Indexierung, Datenerfassung und -analyse, News Feeds usw.
- Manuell / Automatisiert
- Screen Scraping / Web Harvesting / Web Data Extraction



Mein Projekt

- BeautifulSoup und Selenium basierter Scraper
- Unterstützt 3 Sprachen
- Benutzer definiert Standort
- Sammelt Daten über Hotels
- Sortiermöglichkeiten
- Mehrere Möglichkeiten zur Speicherung der erfassten Daten



Warum wird Scraping blockiert?

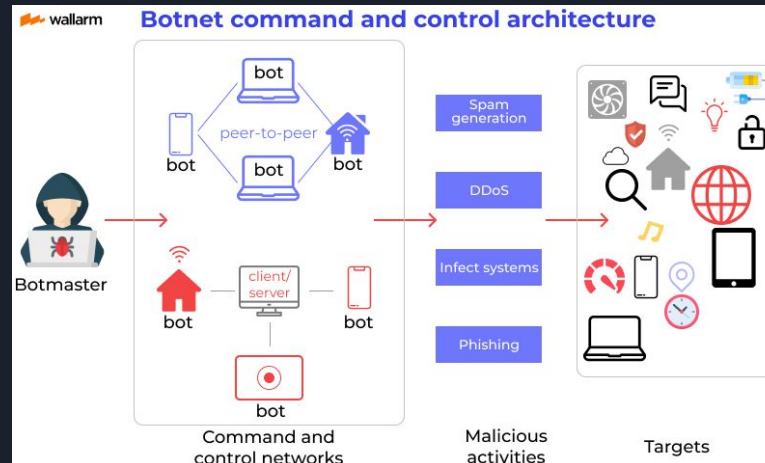
Legitime VS Böartige Scraper

Legitime Scraper

- Teil einer Organisation
- Beachtet die Datei robot.txt
- Verlässt sich auf vertrauenswürdige Server

Böartige Scraper

- Verwendet gefälschte Benutzeragenten
- Ignoriert die Datei robot.txt komplett
- Verlässt sich auf ein Botnet



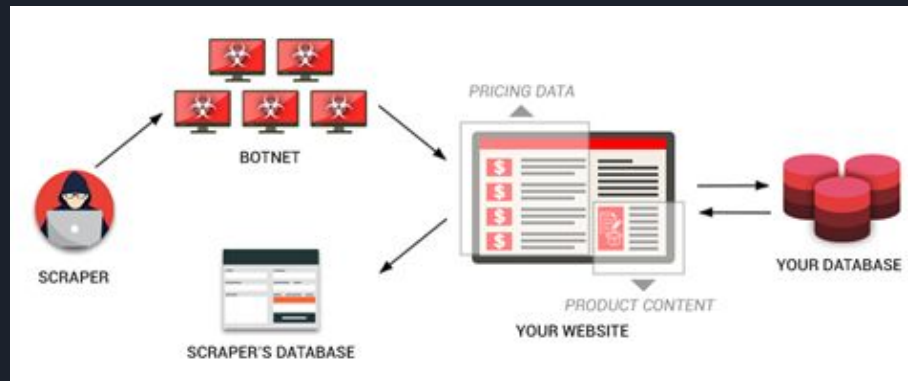
Wie kann Scraping eine Website beschädigen?

Price Scraping

- Ziel: Preisinformationen
- Opfer: Reisebüros, Ticketverkäufer, Online-Elektronikhändler
- Motiv: Einkommen steigern, Kunden gewinnen

Content Scraping

- Ziel: Digitale Inhalte
- Opfer: Online-Produktkataloge, stark von digitalen Inhalten abhängige Unternehmen
- Motiv: Konkurrenzunternehmen beschädigen



Wie blockieren Webseiten Scraper?

- Überwachung neuer oder bestehender Benutzerkonten mit hoher Aktivität und ohne Einkäufe
- Erkennung ungewöhnlich hoher Mengen an Produktaufrufen
- Verfolgung der Aktivität von Wettbewerbern
- Durchsetzung der Website-Nutzungsbedingungen
- Einsatz von Bot-Präventionssoftware mit umfassender Verhaltensanalyse



Vielen Dank fürs
Zuhören! Haben Sie
fragen?

