

# Project 3: Web Scraper with Proxy Rotation

K. Sandeep kumar

# Introduction

- This project builds a web scraper that uses proxy rotation to avoid detection.
- It can solve CAPTCHAs, retry failed requests, and extract structured data.
- Helps in bypassing geo-restrictions and anti-bot systems.

# Features

- Rotating proxy IPs with failover
- Automatic CAPTCHA solving
- Retry system with exponential backoff
- Extract structured data to JSON/CSV
- Mimics human browsing to avoid detection

# Proxy Management

- Sources and validates proxies (HTTP/HTTPS/SOCKS)
- Rotates proxies automatically
- Tests proxy health (speed, success rate, anonymity)
- Removes dead proxies automatically

# HTTP Handling & CAPTCHA

- Requests sent through proxies
- User-Agent rotated with each request
- Cookies and sessions handled correctly
- Detects and solves CAPTCHAs via 2Captcha API







# Data Extraction

- Parses HTML with BeautifulSoup
- Uses CSS selectors or XPath for accuracy
- Supports dynamic content  
(Selenium/Playwright)
- Stores data in JSON or CSV formats

# Resilience & Stealth

- • Retries with exponential backoff
- • Blacklists failing proxies
- • Randomized request intervals
- • Mimics human browsing patterns
- • Avoids honeypot traps

# Skills Gained

-  Advanced HTTP handling
-  Proxy management
-  CAPTCHA solving integration
-  Robust HTML parsing
-  Resilience engineering
-  Operational security & compliance



# Learning Outcomes

- Build enterprise-grade scraping systems
- Master anti-detection techniques
- Integrate third-party APIs
- Design fault-tolerant systems
- Handle real-world scraping challenges
- Develop secure scraping tools