Web Crawling Prevention Measures

Difficulties

- ## Captcha

  Solution: Use capthca api such as https://2captcha.com/ to by pass capthca.

- ## Cookies

  Solution: Developer can get the cookies by requests/selenium libraries.

- ## Limit the number of request

  Solution: Proxies - Rotate proxies before requesting. We can buy paid proxies from websites like sslproxies.org

Prevention measures

All the difficulties which web crawlers are facing, we can use that as a solution for us.

- Captcha
- Cookies
- Limit the number of request
- Anti-Scraping Tools

Anti-Scraping Tools

Today, web developers have tools such as ScrapeSheild and ScrapeSentry that can differentiate bots from humans. Using such tools, web developers can manipulate content shown to bots and humans, and also restrict bots from scraping the website.

Anti-Scraping Tool Example

Cloudflare ScrapeShield

ScrapeShield is a feature in the Cloudflare dashboard that protects your site's content.

- E-mail obfuscation: Scramble e-mail addresses for bots to prevent email spam
- Hotlink protection: Prevent bandwidth abuse
- Server side excludes: hide specific content on your site from suspicious visitors

Conclusion

Making our website impossible to scrap is nearly impossible because websites like amazon, twitter, aajtak, etc. are unable to secure their data from being scraped.

But, we can increase the security to the maximum level like others by using the prevention measures which were discussed. This will help us to detect bots and also secure our website from cyber attacks.