

Forum:[Hack Forums](#)

Thread: [Scraping websites == how to not scrape popupwindow with privacy/cookies/consent?](#)

myplatesarefat - November 23, 2018, 1:15 pm

I cringe because I have made overall good bot that with regex scrapes parts of websites that i need.

But sometimes that consent window is so freaking long it contains 10-30 sentences (more than giiven page sometimes) it just pop ups and asks for ur consent.

i have no idea how to exclude it from scraping.

do they have anything in common in source code? (i mean if popup consent window in one website has anything in common with consent window in another website) because I could not scrape something inside div with ID== soomething, but afaik theres no common thing in source code.

Disease - November 23, 2018, 2:09 pm

(Nov 23, 2018 02:15 PM)myplatesarefat Wrote: [CITE](#) I cringe because I have made overall good bot that with regex scrapes parts of websites that i need.

But sometimes that consent window is so freaking long it contains 10-30 sentences (more than giiven page sometimes) it just pop ups and asks for ur consent.

i have no idea how to exclude it from scraping.

do they have anything in common in source code? (i mean if popup consent window in one website has anything in common with consent window in another website) because I could not scrape something inside div with ID== soomething, but afaik theres no common thing in source code.

Have you thought to look at the source of some of these disclaimers? You're the one parsing them, not us. How would we be able to tell you whether there's a common theme or structure? Depending on what you're scraping, there's a solid chance they're using a common provider or library to handle this. But each of them could be custom written implementations, in which case you're unlikely to find any solid commonalities.

You're probably going to have to target the language used. Rather than focus on HTML structure, look at the wording used. These disclaimers, if I'm right in thinking about what you're referring to, are spurred by EU's GDPR. As such, they may use fairly consistent wording.