The whole Trust+ / Internet Sehat blocklist database, now in one regular expression;

reinhart1010.id – 7 January 2023

(#_)!

At the end of 2022, I decided to experiment on building a *lightweight* Indonesian internet blocklist database, which can be consumed offline.

No network connections to servers of Kominfo, Telkom Indonesia, and community-run services like indi.wtf. Because all you need is a freakin' huge regular expression.

Research methods

We wrote a simple Go script to compile the official Indonesian internet blocklist, found on https://trustpositif.kominfo.go.id, and convert it into a freakin' huge trie. Then that trie is then converted into regular expressions.

And to test whether the regex is effective, we decided to test the generated regex back against the original list of blocked domains.

Results

The experiment grew a 20MB-ish regex file, representing the freakin' huge trie I have mentioned earlier. That said, there's always many ways to improve, including reversing the original domain's arrangement of characters (e.g. "alterine0101.id" □□ "di.1010eniretla") to yield more compact results (because there are more domains ending with ".com" instead of those starting with "www.").

Unfortunately, these gigantic regex files cannot be parsed by Go's own **regexp** system library, hence we decided to use the regexp2 library instead, which is based on Microsoft's regex parses implementation for .NET.

And even if I switch to regexp2, only the reversed version of the regex would work well. I feel confident that the generated regex is 99.9% accurate, tested on Reinhart's M1 MacBook Air with no issues.

You can see my GitHub repo here for the code and the results. Feel free to use that as a benchmark tool for PCRE regex engines out there. We may eventually update the blocked domains list, eventually, to ensure the freshness of these regex-based blocklists.

That's all and (#_)!