

Webscraper by Moop Studios

Webscrapping Process

1. Visual inspection: Figure out what to extract
2. Make HTTP request to page
3. Parse HTTP response
4. Persist/Utilize the relevant data

Links

Google query
syntax: <https://stackoverflow.com/questions/38619478/google-search-web-scraping-with-python>
Elite proxy list: <http://www.gatherproxy.com/proxylist/anonymity/?t=Elite>
Advanced
Webscrapping: <https://www.codementor.io/blog/python-web-scraping-63l2v9sf2q>
Basic webscrapping: <https://towardsdatascience.com/web-scraping-with-python-a-to-z-277a445d64c7>

Obstacles

- Authentication
 - Hidden values (Username/Password)
 - Setting headers
 - Possible HTTP status codes: 401, 403, 407
- Server blacklisting
 - Request rate analyzation
 - Header inspection
 - Honeypots
 - Pattern detection
- Redirects/Captchas
- Infinite scrolling (Facebook, Instagram, etc.)

Ethics

The best rule to follow here is the Golden Rule: do unto others as you would have them do unto you. If a website owner puts anti-webscrapping on their server, it is best leave that website alone.

Overall process



