Webscraper by **Moop Studios**

Webscraping 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

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.)

Links

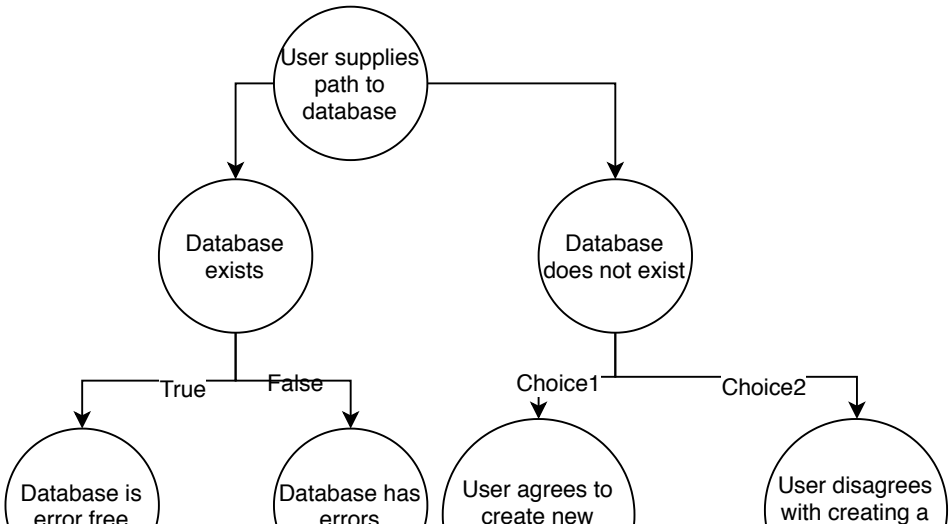
Google query syntax: https://stackoverflow.com/questions/38619478/google-search-web-scraping-with-python

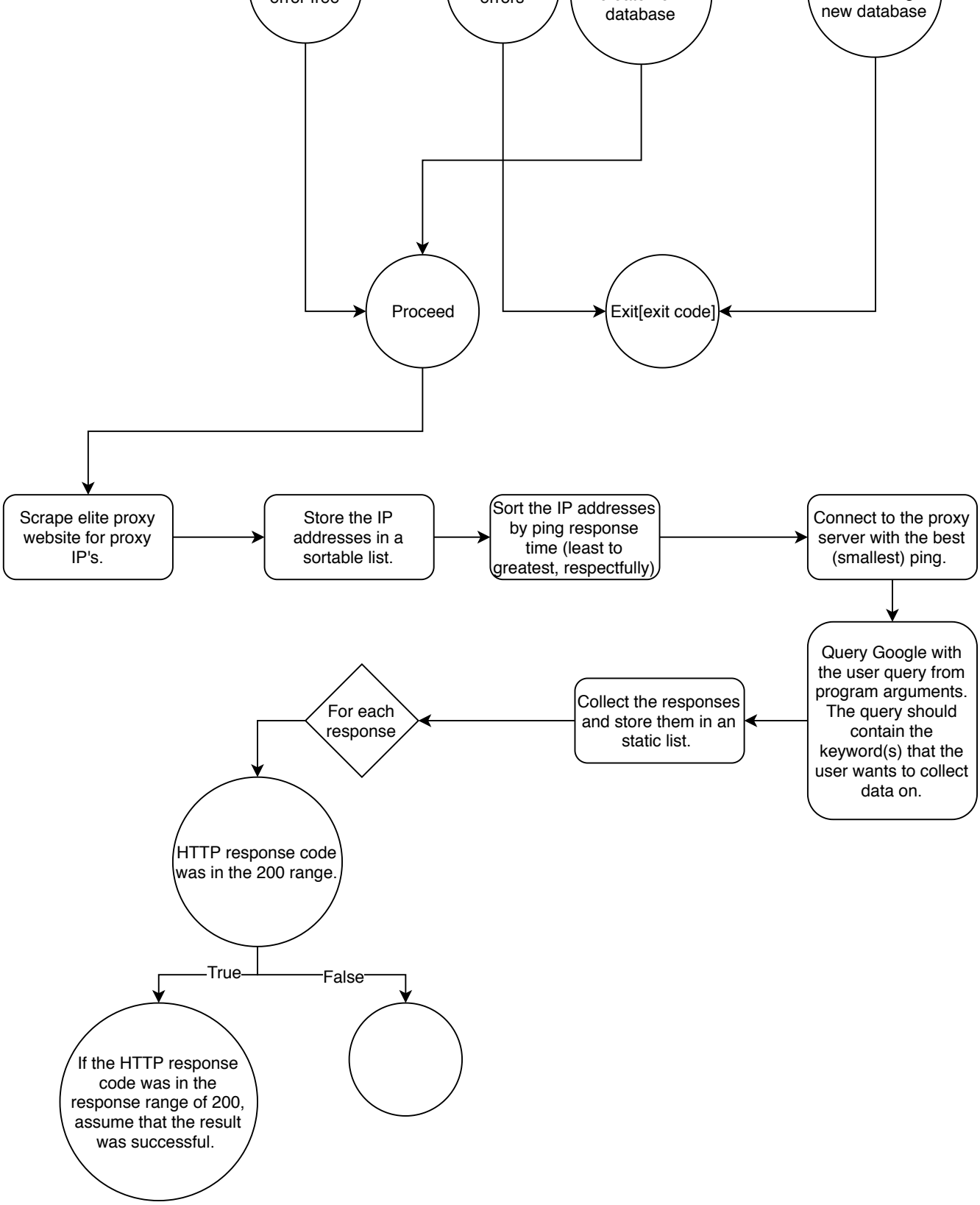
Elite proxy list: http://www.gatherproxy.com/proxylist/anonymity/?t=Elite

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-webscraping on their server, it is best leave that website alone.

Overall process




database
new database
Proceed
Exit[exit code]
Scrape elite proxy website for proxy IP's.
Store the IP addresses in a sortable list.
Sort the IP addresses by ping response time (least to greatest, respectfully)
Connect to the proxy server with the best (smallest) ping.
Query Google with the user query from program arguments. The query should contain the keyword(s) that the user wants to collect data on.
Collect the responses and store them in an static list.
For each response
HTTP response code was in the 200 range.
True
False
If the HTTP response code was in the response range of 200, assume that the result was successful.