FetchOptions

bin/nutch fetch

The fetcher logs to stderr with fetcher output codes.

called java class

net.nutch.fetcher.RequestScheduler

command line options

bin/nutch fetch [-verbose] <dir>

-verbose

config file options

http.agent.name

Our HTTP 'User-Agent' request header.

http.robots.agents

The agent strings we'll look for in robots.txt files, comma-separated, in decreasing order of precedence.

http.agent.description

Further description of our bot- this text is used in the User-Agent header. It appears in parenthesis after the agent name.

http.agent.url

A URL to advertise in the User-Agent header. This will appear in parenthesis after the agent name.

http.agent.email

An email address to advertise in the HTTP 'From' request header and User-Agent header.

http.agent.version

A version string to advertise in the User-Agent header.

http.timeout

The default network timeout, in milliseconds.

http.content.limit

The default length limit for downloaded content, in bytes. Content longer than this is truncated.

http.version.1.1

If true, the fetcher will attempt to use HTTP version 1.1 and gzip encoding.

fetcher.server.delay

The number of seconds the fetcher will delay between successive requests to the same server.

fetcher.threads.fetch

The number of FetcherThreads the fetcher should use. This is also determines the maximum number of requests that are made at once (each FetcherThread handles one connection).

fetcher.threads.output

The number of OutputThreads to use. When adjusting this, remember that each thread could be holding a raw page, it's DOM structure, plaintext, and extracted links in memory.

fetcher.stats.minutes

Controls how often the fetcher will dump progress statistics to the logs, in minutes.

fetcher.request.queue

The maximum number of unfetched requests to queue in memory.

fetcher.output.queue

The maximum number of completed (but unwritten) requests to queue in memory before throttling the fetcher.

fetcher.active.servers

The maximum number of distinct servers that may be referenced by queued requests.

fetcher.robots.cache

The minimum number of robots.txt files to cache for inactive servers.

fetcher.server.maxurls

The maximum number of URLs that may be queued at once for a single host.

fetcher.lowservers.threshold

When there are fewer than this many servers in the fetcher's active queues, each server's queue of URLs will be pruned to fetcher.lowservers.maxurls.

fetcher.lowservers.maxurls

See description of fetcher.lowservers.threshold.

fetcher.retry.max

The maximum number of times the fetcher will attempt to get a page that has encountered recoverable errors.

fetcher.redirect.max

The maximum number of redirects the fetcher will follow when trying to fetch a page.

fetcher.host.consecutive.failures

The maximum number of consecutive failures, excluding 404 errors, to allow on a given server before declaring it dead (note: each failure will have had up to fetcher.retry.max retries).

fetcher.host.max.failerr.rate

The maximum fetch error rate, excluding 404s, to allow for a given server before declaring it dead. Note: errors include transient issues, and multiple retries contribute to the score (so, getting the first page on the 3rd try gives you a .66 "failerr.rate").

fetcher.host.min.requests.rate

A threshold on the minimum number of requests we issue to a host before applying fetcher.host.max.failerr.rate. At least this many requests will be issued before declaring a host dead due to error rate. Note: this setting does not affect fetcher.host.consecutive.failures!

excludehosts.suffix.file

Filename which contains list of hostnames we shouldn't fetch from.

fetcher.trace.longmsg

Whether to use "long messages" is the trace portion of the logged output (if set to false, terse messages will be used).

fetcher.trace.success

Whether to log successful fetches in the trace log.

fetcher.trace.not.found

Whether to log 404/Not Found errors in the trace log.

fetcher.throttle.period

How often throttling behavior should be readjusted based on current bandwidth usage, measured in seconds. Set to -1 to disable throttling.

fetcher.throttle.bandwidth

The desired amount of bandwidth the fetcher should use (aside from DNS and TCP overhead), in kbits/s. Set to -1 to disable throttling. Note: This is **not** a cap, this is a target for bandwidth usage over time.

fetcher.throttle.initial.threads

The number of threads that should be active initially.

– MatthiasJaekle - 13 Mar 2004