scrapy-zyte-smartproxy Documentation

Release 2.2.0

Zyte

Aug 05, 2022

CONFIGURATION

1 Configuration **3**

2 How to use it **5**
2.1 Settings 5

3 All the rest **9**
3.1 Changes 9

scrapy-zyte-smartproxy is a Scrapy downloader middleware to interact with Zyte Smart Proxy Manager (formerly Crawlera) automatically.

CHAPTER

ONE

CONFIGURATION

- Add the Zyte Smart Proxy Manager middleware including it into the `DOWNLOADER_MIDDLEWARES` in your `settings.py` file:

```
DOWNLOADER_MIDDLEWARES = {
    ...
    'scrapy_zyte_smartproxy.ZyteSmartProxyMiddleware': 610
}
```

- Then there are two ways to enable it
  - Through `settings.py`:

```
ZYTE_SMARTPROXY_ENABLED = True
ZYTE_SMARTPROXY_APIKEY = 'apikey'
```

  - Through spider attributes:

```
class MySpider:
    zyte_smartproxy_enabled = True
    zyte_smartproxy_apikey = 'apikey'
```

- (optional) If you are not using the default Zyte Smart Proxy Manager proxy (`http://proxy.zyte.com:8011`), for example if you have a dedicated or private instance, make sure to also set `ZYTE_SMARTPROXY_URL` in `settings.py`, e.g.:

```
ZYTE_SMARTPROXY_URL = 'http://myinstance.zyte.com:8011'
```

CHAPTER

TWO

HOW TO USE IT

2.1 Settings

This Scrapy downloader middleware adds some settings to configure how to work with Zyte Smart Proxy Manager.

2.1.1 ZYTE_SMARTPROXY_APIKEY

Default: `None`

Unique Zyte Smart Proxy Manager API key provided for authentication.

2.1.2 ZYTE_SMARTPROXY_URL

Default: `'http://proxy.zyte.com:8011'`

Zyte Smart Proxy Manager instance URL, it varies depending on adquiring a private or dedicated instance. If Zyte Smart Proxy Manager didn't provide you with a private instance URL, you don't need to specify it.

2.1.3 ZYTE_SMARTPROXY_MAXBANS

Default: `400`

Number of consecutive bans from Zyte Smart Proxy Manager necessary to stop the spider.

2.1.4 ZYTE_SMARTPROXY_DOWNLOAD_TIMEOUT

Default: `190`

Timeout for processing Zyte Smart Proxy Manager requests. It overrides Scrapy's `DOWNLOAD_TIMEOUT`.

2.1.5 ZYTE_SMARTPROXY_PRESERVE_DELAY

Default: `False`

If `False` Sets Scrapy's `DOWNLOAD_DELAY` to `0`, making the spider to crawl faster. If set to `True`, it will respect the provided `DOWNLOAD_DELAY` from Scrapy.

2.1.6 ZYTE_SMARTPROXY_DEFAULT_HEADERS

Default: `{}`

Default headers added only to Zyte Smart Proxy Manager requests. Headers defined on `DEFAULT_REQUEST_HEADERS` will take precedence as long as the `ZyteSmartProxyMiddleware` is placed after the `DefaultHeadersMiddleware`. Headers set on the requests have precedence over the two settings.

- This is the default behavior, `DefaultHeadersMiddleware` default priority is `400` and we recommend `ZyteSmartProxyMiddleware` priority to be `610`

2.1.7 ZYTE_SMARTPROXY_BACKOFF_STEP

Default: `15`

Step size used for calculating exponential backoff according to the formula: `random.uniform(0, min(max, step * 2 ** attempt))`.

2.1.8 ZYTE_SMARTPROXY_BACKOFF_MAX

Default: `180`

Max value for exponential backoff as showed in the formula above.

2.1.9 ZYTE_SMARTPROXY_FORCE_ENABLE_ON_HTTP_CODES

Default: `[]`

List of HTTP response status codes that warrant enabling Zyte Smart Proxy Manager for the corresponding domain.

When a response with one of these HTTP status codes is received after a request that did not go through Zyte Smart Proxy Manager, the request is retried with Zyte Smart Proxy Manager, and any new request to the same domain is also sent through Zyte Smart Proxy Manager.

Settings
: All configurable Scrapy Settings added by the Middleware.

With the middleware, the usage of Zyte Smart Proxy Manager is automatic, every request will go through Zyte Smart Proxy Manager without nothing to worry about. If you want to *disable* Zyte Smart Proxy Manager on a specific Request, you can do so by updating *meta* with *dont_proxy=True*:

```
scrapy.Request(
    'http://example.com',
    meta={
        'dont_proxy': True,
        ...
    },
)
```

Remember that you are now making requests to Zyte Smart Proxy Manager, and the Zyte Smart Proxy Manager service will be the one actually making the requests to the different sites.

If you need to specify special Zyte Smart Proxy Manager headers, just apply them as normal Scrapy headers.

Here we have an example of specifying a Zyte Smart Proxy Manager header into a Scrapy request:

```
scrapy.Request(
    'http://example.com',
    headers={
        'X-Crawlera-Max-Retries': 1,
        ...
    },
)
```

Remember that you could also set which headers to use by default by all requests with DEFAULT_REQUEST_HEADERS

Note: Zyte Smart Proxy Manager headers are removed from requests when the middleware is activated but Zyte Smart Proxy Manager is disabled. For example, if you accidentally disable Zyte Smart Proxy Manager via `zyte_smartproxy_enabled = False` but keep sending `X-Crawlera-*` headers in your requests, those will be removed from the request headers.

This Middleware also adds some configurable Scrapy Settings, check *the complete list here*.

CHAPTER

THREE

ALL THE REST

3.1 Changes

3.1.1 v2.2.0 (2022-08-05)

Added support for Scrapy 2.6.2 and later.

Scrapy 1.4 became the minimum supported Scrapy version.

3.1.2 v2.1.0 (2021-06-16)

- Use a custom logger instead of the root one

3.1.3 v2.0.0 (2021-05-12)

Following the upstream rebranding of Crawlera as Zyte Smart Proxy Manager, `scrapy-crawlera` has been renamed as `scrapy-zyte-smartproxy`, with the following backward-incompatible changes:

- The repository name and Python Package Index (PyPI) name are now `scrapy-zyte-smartproxy`.
- Setting prefixes have switched from `CRAWLERA_` to `ZYTE_SMARTPROXY_`.
- Spider attribute prefixes and request meta key prefixes have switched from `crawlera_` to `zyte_smartproxy_`.
- `scrapy_crawlera` is now `scrapy_zyte_smartproxy`.
- `CrawleraMiddleware` is now `ZyteSmartProxyMiddleware`, and its default `url` is now `http://proxy.zyte.com:8011`.
- Stat prefixes have switched from `crawlera/` to `zyte_smartproxy/`.
- The online documentation is moving to https://scrapy-zyte-smartproxy.readthedocs.io/

Note: Zyte Smart Proxy Manager headers continue to use the `X-Crawlera-` prefix.

- In addition to that, the `X-Crawlera-Client` header is now automatically included in all requests.

3.1.4 v1.7.2 (2020-12-01)

- Use request.meta than response.meta in the middleware

3.1.5 v1.7.1 (2020-10-22)

- Consider Crawlera response if contains *X-Crawlera-Version* header
- Build the documentation in Travis CI and fail on documentation issues
- Update matrix of tests

3.1.6 v1.7.0 (2020-04-01)

- Added more stats to better understanding the internal states.
- Log warning when using *https://* protocol.
- Add default *http://* protocol in case of none provided, and log warning about it.
- Fix duplicated request when the response is not from crawlera, this was causing an infinite loop of retries when *dont_filter=True*.

3.1.7 v1.6.0 (2019-05-27)

- Enable crawlera on demand by setting `CRAWLERA_FORCE_ENABLE_ON_HTTP_CODES`

3.1.8 v1.5.1 (2019-05-21)

- Remove username and password from settings since it's removed from crawlera.
- Include affected spider in logs.
- Handle situations when crawlera is restarted and reply with 407's for a few minutes by retrying the requests with a exponential backoff system.

3.1.9 v1.5.0 (2019-01-23)

- Correctly check for bans in crawlera (Jobs will not get banned on non ban 503's).
- Exponential backoff when crawlera doesn't have proxies available.
- Fix `dont_proxy=False` header disabling crawlera when it is enabled.

3.1.10 v1.4.0 (2018-09-20)

- Remove X-Crawlera-* headers when Crawlera is disabled.
- Introduction of DEFAULT_CRAWLERA_HEADERS settings.

3.1.11 v1.3.0 (2018-01-10)

- Use CONNECT method to contact Crawlera proxy.

3.1.12 v1.2.4 (2017-07-04)

- Trigger PYPI deployments after changes made to TOXENV in v1.2.3

3.1.13 v1.2.3 (2017-06-29)

- Multiple documentation fixes
- Test scrapy-crawlera on combinations of software used by scrapinghub stacks

3.1.14 v1.2.2 (2017-01-19)

- Fix Crawlera error stats key in Python 3.
- Add support for Python 3.6.

3.1.15 v1.2.1 (2016-10-17)

- Fix release date in README.

3.1.16 v1.2.0 (2016-10-17)

- Recommend middleware order to be `610` to run before `RedirectMiddleware`.
- Change default download timeout to 190s or 3 minutes 10 seconds (instead of 1800s or 30 minutes).
- Test and advertize Python 3 compatiblity.
- New `crawlera/request` and `crawlera/request/method/*` stats counts.
- Clear Scrapy DNS cache for proxy URL in case of connection errors.
- Distribute plugin as universal wheel.

Changes
See what has changed in recent scrapy-zyte-smartproxy versions.