darc

Release 0.9.2

Jarry Shaw

Jan 04, 2021

CONTENTS

1 How to . . . 3
- 1.1 How to gracefully deploy `darc`? . . . 3
- 1.2 How to implement a sites customisation? . . . 6
- 1.3 How to implement a custom proxy middleware? . . . 10

2 Technical Documentation 13
- 2.1 Proxy Utilities . . . 15
- 2.2 Sites Customisation . . . 22
- 2.3 Module Constants . . . 23
- 2.4 Custom Exceptions . . . 28
- 2.5 Data Models . . . 31

3 Configuration 33
- 3.1 General Configurations . . . 33
- 3.2 Data Storage . . . 35
- 3.3 Web Crawlers . . . 36
- 3.4 White / Black Lists . . . 38
- 3.5 Data Submission . . . 39
- 3.6 Tor Proxy Configuration . . . 40
- 3.7 I2P Proxy Configuration . . . 41
- 3.8 ZeroNet Proxy Configuration . . . 42
- 3.9 Freenet Proxy Configuration . . . 43

4 Customisations 45
- 4.1 Hooks between Rounds . . . 45
- 4.2 Custom Proxy . . . 46
- 4.3 Sites Customisation . . . 47

5 Docker Integration 51

6 Web Backend Demo 63

7 Data Models Demo 69

8 Submission Data Schema 73
- 8.1 New Host Submission . . . 73
- 8.2 Requests Submission . . . 77
- 8.3 Selenium Submission . . . 82
- 8.4 Model Definitions . . . 85

9 Auxiliary Scripts 91

9.1 Health Check . . . 91
9.2 Upload API Submission Files . . . 91
9.3 Remove Repeated Lines . . . 92
9.4 Redis Clinic . . . 92

10 Rationale **95**

11 Installation **97**

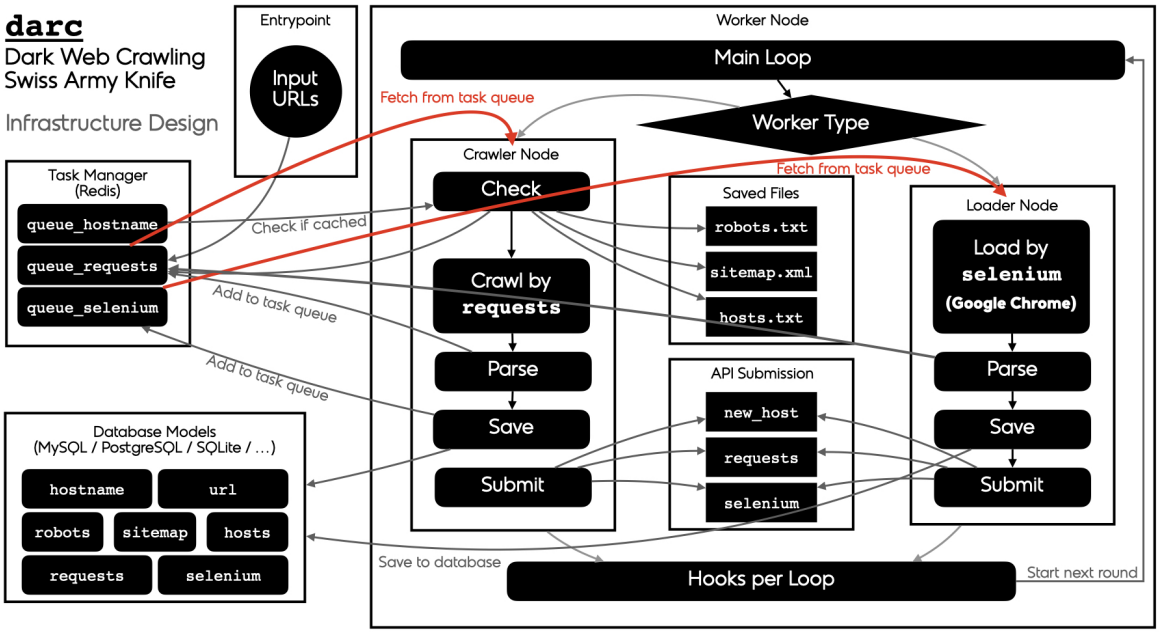
12 Usage **99**

13 Indices and tables **101**

Python Module Index **103**

Index **105**

darc is designed as a swiss army knife for darkweb crawling. It integrates `requests` to collect HTTP request and response information, such as cookies, header fields, etc. It also bundles `selenium` to provide a fully rendered web page and screenshot of such view.



CHAPTER

ONE

HOW TO . . .

This is the knowledge base for *darc* project. Should you request for more articles, please create an issue at the GitHub repository.

1.1 How to gracefully deploy `darc`?

Important: It is **NOT** necessary to work at the `darc` repository folder directly. You can just use `darc` with your customised code somewhere as you wish.

However, for simplicity, all relative paths referred in this article is relative to the project root of the `darc` repository.

To deploy `darc`, there would generally be three basic steps:

1. deploy the `darc` Docker image;
2. setup the `healthcheck` watchdog service;
3. install the `upload` cron job (optional)

1.1.1 To Start With

Per best practice, the system must have as least **2 GB RAM** and **2 CPU cores** to handle the `loader` worker properly. And the capacity of the RAM will heavily impact the performance of the `selenium` integration as Google Chrome is the renowned memory monster.

Note: Imma assume that you're using *NIX systems, as I don't believe a Windows user is gonna see this ;)

Firstly, you will need to clone the repository to your system:

```
git clone https://github.com/JarryShaw/darc.git
# change your working directory
cd darc
```

then set up the folders you need for the log files:

```
mkdir -p logs
mkdir -p logs/cron
```

And now, you will need to decide where you would like to store the data (documents crawled and saved by `darc`); let's assume that you have a `/data` disk mounted on the system – since that's what I have on mine xD – which would be big enough to use as a safe seperated storage place from the system so that `darc` will not crash your system by exhausting the storage,

```
mkdir /data/darc
# and make a shortcut
ln -s /data/darc data
```

therefore, you're gonna save your data in `/data/darc` folder.

1.1.2 Software Dependency

After setting local systems, there're some software dependencies you shall install:

1. Docker

`darc` is exclusively deployed through Docker environment, even though it can also be deployed directly on a host machine, either Linux or macOS, and perhaps Windows but I had never tested.

2. Database

`darc` needs database backend for the task queue management and other stuffs. It is highly recommended to deploy `darc` with Redis; but if you insist, you may use relationship database (e.g. MySQL, SQLite, PostgreSQL, etc.) instead.

Important: In this article, I will not discuss about the usage of relationship database as they're just too miserable for `darc` in terms of availability anyway.

As per best practice, *4 GB RAM* would be minimal requirement for the Redis database. It would be suggested to use directly a cloud provider hosted Redis database instead of running it on the same server as `darc`.

1.1.3 Deploy `darc` Docker Image

As discussed in *Docker Integration*, `darc` is exclusively integrated with Docker workflow. So basically, just pull the image from Docker Hub or GitHub Container Registry:

```
# Docker Hub
docker pull jsnbzh/darc:latest
# GitHub Container Registry
docker pull ghcr.io/jarryshaw/darc:latest
```

In cases where you would like to use a *debug* image, which changes the `apt` sources to China hosted and IPython and other auxiliaries installed, you call also pull such image instead:

```
# Docker Hub
docker pull jsnbzh/darc:debug
# GitHub Container Registry
docker pull ghcr.io/jarryshaw/darc-debug:latest
```

Then you will need to customise the `docker-compose.yml` based on your needs. Default values and descriptive help messages can be found in the file.

The rest of it is easy as just calling `docker-compose` command to manage the deployed containers, thus I shall not discuss further.

Deploy with Customisations

Important: I've made a sample customisation at `demo/deploy/` folder, which can be used directly as a new repository to start with your customisation, please check it out before moving forwards.

As in the sample customisation, you can simply use the `Dockerfile` there as your Docker environment declration. And the entrypoint file `market/run.py` has the sites customisations registered and the CLI bundled.

1.1.4 Setup `healthcheck` Daemon Service

Since `darc` can be quite a burden to its host system, I introduced this healthcheck service as discussed in *Auxiliary Scripts*.

For a normal **System V** based service system, you can simply install the `darc-healthcheck` service to `/etc/systemd/system/`:

```
ln -s extra/healthcheck.service /etc/systemd/system/darc-healthcheck.service
```

then enable it to run at startup:

```
sudo systemctl enable darc-healthcheck.service
```

And from now on, you can simply manage the `darc-healthcheck` service through `systemctl` or `service` command as you prefer.

1.1.5 Install `upload` Cron Job

In certain cases, you might wish to upload the API submission JSON files to your FTP server which has much more space than the deploy server, then you can utilise the `upload` cron job as mentioned in *Auxiliary Scripts*.

Simply type the following command:

```
crontab -e
```

and add the cron job into the file opened:

```
10 0 * * * ( cd /path/to/darc/ && /path/to/python3 /path/to/darc/extra/upload.py --
↪host ftp://hostname --user username:password ) >> /path/to/darc/logs/cron/darc-
↪upload.log 2>&1
```

just remember to change the paths, hostname and credential respectively; and at last, to activate the new cron job:

```
sudo systemctl restart cron.service
```

Now, `darc` API submission JSON files will be uploaded to the target FTP server everyday at *0:10 am*.

1.1.6 Bonus Tip

There is a `Makefile` at the project root. You can play and try to exploit it. A very useful command is that

```
make reload
```

when you wish to pull the remote repository and restart `darc` gracefully.

1.2 How to implement a sites customisation?

As had been discussed already in the documentation, the implementation of a sites customisation is dead simple: just inherits the `darc.sites.BaseSite` class and overwrites the corresponding `crawler()` and `loader()` *abstract* **static methods**.

See below an example from the documentation.

```
from darc.sites import BaseSite, register
```

As the class below suggests, you may implement and register your sites customisation for *mysite.com* and *www.mysite.com* using the `MySite` class, where `hostname` attribute contains the list of hostnames to which the class should be associated with.

NB: Implementation details of the `crawler` and `loader` methods will be discussed in following sections.

```
class MySite(BaseSite):
    """This is a site customisation class for demonstration purpose.
    You may implement a module as well should you prefer."""

    #: List[str]: Hostnames the sites customisation is designed for.
    hostname = ['mysite.com', 'www.mysite.com']

    @staticmethod
    def crawler(timestamp, session, link): ...

    @staticmethod
    def loader(timestamp, driver, link): ...
```

Should your sites customisation be associated with multiple sites, you can just add them all to the `hostname` attribute; when you call `darc.sites.register()` to register your sites customisation, the function will automatically handle the registry association information.

```
# register sites implicitly
register(MySite)
```

Nonetheless, in case where you would rather specify the hostnames at runtime (instead of adding them to the `hostname` attribute), you may just leave out the `hostname` attribute as `None` and specify your list of hostnames at `darc.sites.register()` function call.

```
# register sites explicitly
register(MySite, 'mysite.com', 'www.mysite.com')
```

1.2.1 Crawler Hook

The `crawler` method is based on `requests.Session` objects and returns a `requests.Response` instance representing the *crawled* web page.

Type annotations of the method can be described as

```
@staticmethod
def crawler(session: requests.Session, link: darc.link.Link) -> requests.Response: ...
```

where `session` is the `requests.Session` instance with **proxy** presets and `link` is the target link (parsed by `darc.link.parse_link()` to provide more information than mere string).

For example, let's say you would like to inject a cookie named `SessionID` and an `Authentication` header with some fake identity, then you may write the `crawler` method as below.

```
@staticmethod
def crawler(timestamp, session, link):
    """Crawler hook for my site.

    Args:
        timestamp (datetime.datetime): Timestamp of the worker node reference.
        session (requests.Session): Session object with proxy settings.
        link (darc.link.Link): Link object to be crawled.

    Returns:
        requests.Response: The final response object with crawled data.

    """
    # inject cookies
    session.cookies.set('SessionID', 'fake-session-id-value')

    # insert headers
    session.headers['Authentication'] = 'Basic fake-identity-credential'

    response = session.get(link.url, allow_redirects=True)
    return response
```

In this case when *darc* crawling the link, the HTTP(S) request will be provided with a session cookie and HTTP header, so that it may bypass potential authorisation checks and land on the target page.

1.2.2 Loader Hook

The `loader` method is based on `selenium.webdriver.Chrome` objects and returns a the original web driver instance containing the *loaded* web page.

Type annotations of the method can be described as

```
@staticmethod
def loader(driver: selenium.webdriver.Chrome, link: darc.link.Link) -> selenium.
↪webdriver.Chrome: ...
```

where `driver` is the `selenium.webdriver.Chrome` instance with **proxy** presets and `link` is the target link (parsed by `darc.link.parse_link()` to provide more information than mere string).

For example, let's say you would like to animate user login and go to the target page after successful attempt, then you may write the `loader` method as below.

```python
@staticmethod
def loader(timestamp, driver, link):
    """Loader hook for my site.

    Args:
        timestamp: Timestamp of the worker node reference.
        driver (selenium.webdriver.Chrome): Web driver object with proxy settings.
        link (darc.link.Link): Link object to be loaded.

    Returns:
        selenium.webdriver.Chrome: The web driver object with loaded data.

    """
    # land on login page
    driver.get('https://%s/login' % link.host)

    # animate login attempt
    form = driver.find_element_by_id('login-form')
    form.find_element_by_id('username').send_keys('admin')
    form.find_element_by_id('password').send_keys('p@ssd')
    form.click()

    # check if the attempt succeeded
    if driver.title == 'Please login!':
        raise ValueError('failed to login %s' % link.host)

    # go to the target page
    driver.get(link.url)

    # wait for page to finish loading
    from darc.const import SE_WAIT  # should've been put with the top-level import␣
↪statements
    if SE_WAIT is not None:
        time.sleep(SE_WAIT)

    return driver
```

In this case when *darc* loading the link, the web driver will first perform user login, so that it may bypass potential authorisation checks and land on the target page.

1.2.3 In case to drop the link from task queue...

In some scenarios, you may want to remove the target link from the task queue, then there're basically two ways:

1. do like a wildling, remove it directly from the database

As there're three task queues used in *darc*, each represents task queues for the *crawler* (`requests` database) and *loader* (`selenium` database) worker nodes and a track record for known hostnames (hostname database), you will need to call corresponding functions to remove the target link from the database desired.

Possible functions are as below:

- `darc.db.drop_hostname()`
- `darc.db.drop_requests()`
- `darc.db.drop_selenium()`

all take one positional argument `link`, i.e. the `darc.link.Link` object to be removed.

Say you would like to remove `https://www.mysite.com` from the `requests` database, then you may just run

```
from darc.db import drop_requests
from darc.link import parse_link

link = parse_link('https://www.mysite.com')
drop_requests(link)
```

2. or make it in an elegant way

When implementing the sites customisation, you may wish to drop certain links at runtime, then you may simply raise *darc.error.LinkNoReturn* in the corresponding `crawler` and/or `loader` methods.

For instance, you would like to proceed with `mysite.com` but **NOT** `www.mysite.com` in the sites customisation, then you may implement your class as

```
from darc.error import LinkNoReturn

class MySite(BaseSite):

    ...

    @staticmethod
    def crawler(timestamp, session, link):
        if link.host == 'www.mysite.com':
            raise LinkNoReturn(link)

        ...

    @staticmethod
    def loader(timestamp, driver, link):
        if link.host == 'www.mysite.com':
            raise LinkNoReturn(link)

        ...
```

1.2.4 Then what should I do to include my sites customisation?

Simple as well!

Just *install* your codes to where you're running *darc*, e.g. the Docker container, remote server, etc.; then change the startup by injecting your codes before the entrypoint.

Say the structure of the working directory is as below:

```
.
|-- .venv/
|   |-- lib/python3.8/site-packages
|   |   |-- darc/
|   |   |   |-- ...
|   |   |-- ...
|   |-- ...
|-- mysite.py
|-- ...
```

where `.venv` is the folder of virtual environment with *darc* installed and `mysite.py` is the file with your sites customisation.

Then you just need to change your `mysite.py` with some additional lines as below:

```
# mysite.py

import sys

from darc.__main__ import main
from darc.sites import BaseSite, register


class MySite(BaseSite):

    ...


# register sites
register(MySite)

if __name__ == '__main__':
    sys.exit(main())
```

And now, you can start *darc* through `python mysite.py [...]` instead of `python -m darc [...]` with your sites customisation registered to the system.

See also:

`mysite.py`

1.3 How to implement a custom proxy middleware?

As had been discussed already in the documentation, the implementation of a custom proxy is merely two *factory* functions: one yields a `requests.Session` and/or `requests_futures.sessions.FuturesSession` instance, one yields a `selenium.webdriver.Chrome` instance; both with proxy presets.

See below an example from the documentation.

```
from darc.proxy import register
```

1.3.1 Session Factory

The session factory returns a `requests.Session` and/or `requests_futures.sessions.FuturesSession` instance with presets, e.g. proxies, user agent, etc.

Type annocation of the function can be described as

```
def get_session(futures=False) -> requests.Session: ...

@typing.overload
def get_session(futures=True) -> requests_futures.sessions.FuturesSession: ...
```

For example, let's say you're implementing a Socks5 proxy for *localhost:9293*, with other presets same as the default factory function, c.f. `darc.requests.null_session()`.

```
import requests
import requests_futures.sessions

```

(continues on next page)

(continued from previous page)

```
from darc.const import DARC_CPU
from darc.requests import default_user_agent


def socks5_session(futures=False):
    """Socks5 proxy session.

    Args:
        futures: If returns a :class:`requests_futures.FuturesSession`.

    Returns:
        Union[requests.Session, requests_futures.FuturesSession]:
        The session object with Socks5 proxy settings.

    """
    if futures:
        session = requests_futures.sessions.FuturesSession(max_workers=DARC_CPU)
    else:
        session = requests.Session()

    session.headers['User-Agent'] = default_user_agent(proxy='Socks5')
    session.proxies.update({
        'http': 'socks5h://localhost:9293',
        'https': 'socks5h://localhost:9293',
    })
    return session
```

In this case when *darc* needs to use a Socks5 session for its *crawler* worker nodes, it will call the `socks5_session` function to obtain a preset session instance.

1.3.2 Driver Factory

The driver factory returns a `selenium.webdriver.Chrome` instance with presets, e.g. proxies, options/switches, etc.

Type annocation of the function can be described as

```
def get_driver() -> selenium.webdriver.Chrome: ...
```

For example, let's say you're implementing a Socks5 proxy for *localhost:9293*, with other presets same as the default factory function, c.f. `darc.selenium.null_driver()`.

```
import selenium.webdriver
import selenium.webdriver.common.proxy

from darc.selenium import BINARY_LOCATION


def socks5_driver():
    """Socks5 proxy driver.

    Returns:
        selenium.webdriver.Chrome: The web driver object with Socks5 proxy settings.

    """
```

(continues on next page)

(continued from previous page)

```
    options = selenium.webdriver.ChromeOptions()
    options.binary_location = BINARY_LOCATION
    options.add_argument('--proxy-server=socks5://localhost:9293')
    options.add_argument('--host-resolver-rules="MAP * ~NOTFOUND , EXCLUDE localhost"
↪')

    proxy = selenium.webdriver.Proxy()
    proxy.proxyType = selenium.webdriver.common.proxy.ProxyType.MANUAL
    proxy.http_proxy = 'socks5://localhost:9293'
    proxy.ssl_proxy = 'socks5://localhost:9293'

    capabilities = selenium.webdriver.DesiredCapabilities.CHROME.copy()
    proxy.add_to_capabilities(capabilities)

    driver = selenium.webdriver.Chrome(options=options,
                                       desired_capabilities=capabilities)
    return driver
```

In this case when *darc* needs to use a Socks5 driver for its *loader* worker nodes, it will call the `socks5_driver` function to obtain a preset driver instance.

1.3.3 What should I do to register the proxy?

All proxies are managed in the *darc.proxy* module and you can register your own proxy through `darc.proxy.register()`:

```
# register proxy
register('socks5', socks5_session, socks5_driver)
```

As the codes above suggest, the `darc.proxy.register()` takes three positional arguments: proxy type, session and driver factory functions.

See also:

`socks5.py`

CHAPTER

TWO

TECHNICAL DOCUMENTATION

darc is designed as a swiss army knife for darkweb crawling. It integrates `requests` to collect HTTP request and response information, such as cookies, header fields, etc. It also bundles `selenium` to provide a fully rendered web page and screenshot of such view.

`darc.parse.`**`URL_PAT: List[re.Pattern]`**

Regular expression patterns to match all reasonable URLs.

Currently, we have two builtin patterns:

1. HTTP(S) and other *regular* URLs, e.g. WebSocket, IRC, etc.

```
re.compile(r'(?P<url>((https?|wss?|irc):)?(//)?\w+(\.\w+)+/?\S*)', re.UNICODE),
```

2. Bitcoin accounts, data URIs, (ED2K) magnet links, email addresses, telephone numbers, JavaScript functions, etc.

```
re.compile(r'(?P<url>(bitcoin|data|ed2k|magnet|mailto|script|tel):\w+)', re.ASCII)
```

Environ *DARC_URL_PAT*

See also:

The patterns are used in `darc.parse.extract_links_from_text()`.

`darc.save.`**`_SAVE_LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]`**

I/O lock for saving link hash database `link.csv`.

See also:

- `darc.save.save_link()`
- *`darc.const.get_lock()`*

`darc.db.`**`BULK_SIZE: int`**

Default `100`

Environ *DARC_BULK_SIZE*

Bulk size for updating Redis databases.

See also:

- `darc.db.save_requests()`
- `darc.db.save_selenium()`

darc.db.**LOCK_TIMEOUT: Optional[float]**

Default 10

Environ DARC_LOCK_TIMEOUT

Lock blocking timeout.

Note: If is an infinit `inf`, no timeout will be applied.

See also:

Get a lock from `darc.db.get_lock()`.

darc.db.**MAX_POOL: int**

Default 1_000

Environ *DARC_MAX_POOL*

Maximum number of links loading from the database.

Note: If is an infinit `inf`, no limit will be applied.

darc.db.**REDIS_LOCK: bool**

Default False

Environ DARC_REDIS_LOCK

If use Redis (Lua) lock to ensure process/thread-safely operations.

See also:

Toggles the behaviour of `darc.db.get_lock()`.

darc.db.**RETRY_INTERVAL: int**

Default 10

Environ DARC_RETRY

Retry interval between each Redis command failure.

Note: If is an infinit `inf`, no interval will be applied.

See also:

Toggles the behaviour of `darc.db.redis_command()`.

darc.submit.**PATH_API = '{PATH_DB}/api/'**
Path to the API submittsion records, i.e. `api` folder under the root of data storage.

See also:

- *darc.const.PATH_DB*

darc.submit.**SAVE_DB: bool**
Save submitted data to database.

Default True

Environ *SAVE_DB*

darc.submit.**API_RETRY: int**
Retry times for API submission when failure.

Default 3

Environ *API_RETRY*

darc.submit.**API_NEW_HOST: str**
API URL for `submit_new_host()`.

Default None

Environ *API_NEW_HOST*

darc.submit.**API_REQUESTS: str**
API URL for `submit_requests()`.

Default None

Environ *API_REQUESTS*

darc.submit.**API_SELENIUM: str**
API URL for `submit_selenium()`.

Default None

Environ *API_SELENIUM*

Note: If *API_NEW_HOST*, *API_REQUESTS* and *API_SELENIUM* is None, the corresponding submit function will save the JSON data in the path specified by *PATH_API*.

See also:

The *darc* provides a demo on how to implement a *darc*-compliant web backend for the data submission module. See the *demo* page for more information.

darc.selenium.**BINARY_LOCATION: Optional[str]**
Path to Google Chrome binary location.

Default google-chrome

Environ *CHROME_BINARY_LOCATION*

2.1 Proxy Utilities

The *darc.proxy* module provides various proxy support to the *darc* project.

darc.proxy.bitcoin.**PATH = '{PATH_MISC}/bitcoin.txt'**
Path to the data storage of bitcoin addresses.

See also:

- *darc.const.PATH_MISC*

darc.proxy.bitcoin.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext**
I/O lock for saving bitcoin addresses *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.data.**PATH = '{PATH_MISC}/data/'**
Path to the data storage of data URI schemes.

See also:

- *darc.const.PATH_MISC*

darc.proxy.ed2k.**PATH = '{PATH_MISC}/ed2k.txt'**
Path to the data storage of bED2K magnet links.

See also:

- *darc.const.PATH_MISC*

darc.proxy.ed2k.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving ED2K magnet links *PATH*.

See also:

- *darc.const.get_lock()*

The following constants are configuration through environment variables:

darc.proxy.freenet.**FREENET_PORT: int**
Port for Freenet proxy connection.

Default 8888

Environ *FREENET_PORT*

darc.proxy.freenet.**FREENET_RETRY: int**
Retry times for Freenet bootstrap when failure.

Default 3

Environ *FREENET_RETRY*

darc.proxy.freenet.**BS_WAIT: float**
Time after which the attempt to start Freenet is aborted.

Default 90

Environ FREENET_WAIT

Note: If not provided, there will be **NO** timeouts.

---

darc.proxy.freenet.**FREENET_PATH: str**
Path to the Freenet project.

Default /usr/local/src/freenet

Environ *FREENET_PATH*

darc.proxy.freenet.**FREENET_ARGS: List[str]**
Freenet bootstrap arguments for `run.sh start`.

If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

Default ''

Environ *FREENET_ARGS*

Note: The command will be run as *DARC_USER*, if current user (c.f. `getpass.getuser()`) is *root*.

The following constants are defined for internal usage:

`darc.proxy.freenet.`**_MNG_FREENET**`: bool`
If manage Freenet proxy through *darc*.

Default `True`

Environ `DARC_FREENET`

`darc.proxy.freenet.`**_FREENET_BS_FLAG**`: bool`
If the Freenet proxy is bootstrapped.

`darc.proxy.freenet.`**_FREENET_PROC**`: subprocess.Popen`
Freenet proxy process running in the background.

`darc.proxy.freenet.`**_FREENET_ARGS**`: List[str]`
Freenet proxy bootstrap arguments.

`darc.proxy.i2p.`**I2P_REQUESTS_PROXY**`: Dict[str, Any]`
Proxy for I2P sessions.

See also:

- `darc.requests.i2p_session()`

`darc.proxy.i2p.`**I2P_SELENIUM_PROXY**`: selenium.webdriver.common.proxy.Proxy`
`Proxy` for I2P web drivers.

See also:

- `darc.selenium.i2p_driver()`

The following constants are configuration through environment variables:

`darc.proxy.i2p.`**I2P_PORT**`: int`
Port for I2P proxy connection.

Default `4444`

Environ *I2P_PORT*

`darc.proxy.i2p.`**I2P_RETRY**`: int`
Retry times for I2P bootstrap when failure.

Default `3`

Environ *I2P_RETRY*

`darc.proxy.i2p.`**BS_WAIT**`: float`
Time after which the attempt to start I2P is aborted.

Default `90`

Environ `I2P_WAIT`

Note: If not provided, there will be **NO** timeouts.

darc.proxy.i2p.**I2P_ARGS: List[str]**
I2P bootstrap arguments for `i2prouter start`.

If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

Default `''`

Environ *I2P_ARGS*

Note: The command will be run as *DARC_USER*, if current user (c.f. `getpass.getuser()`) is *root*.

The following constants are defined for internal usage:

darc.proxy.i2p.**_MNG_I2P: bool**
If manage I2P proxy through *darc*.

Default `True`

Environ `DARC_I2P`

darc.proxy.i2p.**_I2P_BS_FLAG: bool**
If the I2P proxy is bootstrapped.

darc.proxy.i2p.**_I2P_PROC: subprocess.Popen**
I2P proxy process running in the background.

darc.proxy.i2p.**_I2P_ARGS: List[str]**
I2P proxy bootstrap arguments.

darc.proxy.irc.**PATH = '{PATH_MISC}/irc.txt'**
Path to the data storage of IRC addresses.

See also:

- *darc.const.PATH_MISC*

darc.proxy.irc.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving IRC addresses *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.magnet.**PATH = '{PATH_MISC}/magnet.txt'**
Path to the data storage of magnet links.

See also:

- *darc.const.PATH_MISC*

darc.proxy.magnet.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving magnet links *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.mail.**PATH = '{PATH_MISC}/mail.txt'**
Path to the data storage of email addresses.

See also:

- *darc.const.PATH_MISC*

darc.proxy.mail.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving email addresses *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.null.**PATH = '{PATH_MISC}/invalid.txt'**
Path to the data storage of links with invalid scheme.

See also:

- *darc.const.PATH_MISC*

darc.proxy.null.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving links with invalid scheme *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.script.**PATH = '{PATH_MISC}/script.txt'**
Path to the data storage of bitcoin addresses.

See also:

- *darc.const.PATH_MISC*

darc.proxy.script.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving JavaScript links *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.tel.**PATH = '{PATH_MISC}/tel.txt'**
Path to the data storage of bitcoin addresses.

See also:

- *darc.const.PATH_MISC*

darc.proxy.tel.**LOCK: Union[multiprocessing.Lock, threading.Lock, contextlib.nullcontext]**
I/O lock for saving telephone numbers *PATH*.

See also:

- *darc.const.get_lock()*

darc.proxy.tor.**TOR_REQUESTS_PROXY: Dict[str, Any]**
Proxy for Tor sessions.

See also:

- darc.requests.tor_session()

darc.proxy.tor.**TOR_SELENIUM_PROXY: selenium.webdriver.common.proxy.Proxy**
Proxy for Tor web drivers.

See also:

- darc.selenium.tor_driver()

The following constants are configuration through environment variables:

darc.proxy.tor.**TOR_PORT: int**
Port for Tor proxy connection.

Default 9050

Environ *TOR_PORT*

darc.proxy.tor.**TOR_CTRL: int**
Port for Tor controller connection.

Default 9051

Environ *TOR_CTRL*

darc.proxy.tor.**TOR_PASS: str**
Tor controller authentication token.

Default None

Environ *TOR_PASS*

Note: If not provided, it will be requested at runtime.

---

darc.proxy.tor.**TOR_RETRY: int**
Retry times for Tor bootstrap when failure.

Default 3

Environ *TOR_RETRY*

darc.proxy.tor.**BS_WAIT: float**
Time after which the attempt to start Tor is aborted.

Default 90

Environ TOR_WAIT

Note: If not provided, there will be **NO** timeouts.

---

darc.proxy.tor.**TOR_CFG: Dict[str, Any]**
Tor bootstrap configuration for stem.process.launch_tor_with_config().

Default {}

Environ *TOR_CFG*

Note: If provided, it will be parsed from a JSON encoded string.

The following constants are defined for internal usage:

darc.proxy.tor.**_MNG_TOR: bool**
If manage Tor proxy through *darc*.

Default `True`

Environ `DARC_TOR`

darc.proxy.tor.**_TOR_BS_FLAG: bool**
If the Tor proxy is bootstrapped.

darc.proxy.tor.**_TOR_PROC: subprocess.Popen**
Tor proxy process running in the background.

darc.proxy.tor.**_TOR_CTRL: stem.control.Controller**
Tor controller process (`stem.control.Controller`) running in the background.

darc.proxy.tor.**_TOR_CONFIG: List[str]**
Tor bootstrap configuration for `stem.process.launch_tor_with_config()`.

The following constants are configuration through environment variables:

darc.proxy.zeronet.**ZERONET_PORT: int**
Port for ZeroNet proxy connection.

Default `43110`

Environ *ZERONET_PORT*

darc.proxy.zeronet.**ZERONET_RETRY: int**
Retry times for ZeroNet bootstrap when failure.

Default `3`

Environ *ZERONET_RETRY*

darc.proxy.zeronet.**BS_WAIT: float**
Time after which the attempt to start ZeroNet is aborted.

Default `90`

Environ `ZERONET_WAIT`

Note: If not provided, there will be **NO** timeouts.

darc.proxy.zeronet.**ZERONET_PATH: str**
Path to the ZeroNet project.

Default `/usr/local/src/zeronet`

Environ *ZERONET_PATH*

darc.proxy.zeronet.**ZERONET_ARGS: List[str]**
ZeroNet bootstrap arguments for `run.sh start`.

If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

Default `''`

Environ *ZERONET_ARGS*

Note: The command will be run as *DARC_USER*, if current user (c.f. `getpass.getuser()`) is *root*.

The following constants are defined for internal usage:

`darc.proxy.zeronet._MNG_ZERONET: bool`
If manage ZeroNet proxy through *darc*.

Default `True`

Environ `DARC_ZERONET`

`darc.proxy.zeronet._ZERONET_BS_FLAG: bool`
If the ZeroNet proxy is bootstrapped.

`darc.proxy.zeronet._ZERONET_PROC: subprocess.Popen`
ZeroNet proxy process running in the background.

`darc.proxy.zeronet._ZERONET_ARGS: List[str]`
ZeroNet proxy bootstrap arguments.

To tell the *darc* project which proxy settings to be used for the `requests.Session` objects and `WebDriver` objects, you can specify such information in the *darc.proxy.LINK_MAP* mapping dictionarty.

`darc.proxy.LINK_MAP: DefaultDict[str, Tuple[types.FunctionType, types.FunctionType]]`

```
LINK_MAP = collections.defaultdict(
    lambda: (darc.requests.null_session, darc.selenium.null_driver),
    {
        'tor': (darc.requests.tor_session, darc.selenium.tor_driver),
        'i2p': (darc.requests.i2p_session, darc.selenium.i2p_driver),
    }
)
```

The mapping dictionary for proxy type to its corresponding `requests.Session` factory function and `WebDriver` factory function.

The fallback value is the no proxy `requests.Session` object (`null_session()`) and `WebDriver` object (`null_driver()`).

See also:

- `darc.requests` – `requests.Session` factory functions
- `darc.selenium` – `WebDriver` factory functions

2.2 Sites Customisation

As websites may have authentication requirements, etc., over its content, the *darc.sites* module provides sites customisation hooks to both `requests` and `selenium` crawling processes.

Important: To create a sites customisation, define your class by inheriting `darc.sites.BaseSite` and register it to the *darc* module through `darc.sites.register()`.

To start with, you just need to define your sites customisation by inheriting `BaseSite` and overload corresponding `crawler()` and/or `loader()` methods.

To customise behaviours over `requests`, you sites customisation class should have a `crawler()` method, e.g. `DefaultSite.crawler`.

The function takes the `requests.Session` object with proxy settings and a `Link` object representing the link to be crawled, then returns a `requests.Response` object containing the final data of the crawling process.

To customise behaviours over `selenium`, you sites customisation class should have a `loader()` method, e.g. `DefaultSite.loader`.

The function takes the `WebDriver` object with proxy settings and a `Link` object representing the link to be loaded, then returns the `WebDriver` object containing the final data of the loading process.

To tell the *darc* project which sites customisation module it should use for a certain hostname, you can register such module to the *SITEMAP* mapping dictionary through `register()`:

darc.sites.**SITEMAP: DefaultDict[str, Type[darc.sites._abc.BaseSite]]**

```
from darc.sites.default import DefaultSite

SITEMAP = collections.defaultdict(lambda: DefaultSite, {
    # 'www.sample.com': SampleSite,  # local customised class
})
```

The mapping dictionary for hostname to sites customisation classes.

The fallback value is `darc.sites.default.DefaultSite`.

See also:

Please refer to *Customisations* for more examples and explanations.

2.3 Module Constants

2.3.1 Auxiliary Function

darc.const.**get_lock**()
Get a lock.

Returns Lock context based on *FLAG_MP* and *FLAG_TH*.

Return type Union[ProcessLock, ThreadLock, nullcontext]

2.3.2 General Configurations

darc.const.**REBOOT: bool**
If exit the program after first round, i.e. crawled all links from the `requests` link database and loaded all links from the `selenium` link database.

This can be useful especially when the capacity is limited and you wish to save some space before continuing next round. See *Docker integration* for more information.

Default `False`

Environ *DARC_REBOOT*

darc.const.**DEBUG: bool**
If run the program in debugging mode.

Default `False`

Environ *DARC_DEBUG*

`darc.const.`**VERBOSE: bool**
If run the program in verbose mode. If *DEBUG* is `True`, then the verbose mode will be always enabled.

Default `False`

Environ *DARC_VERBOSE*

`darc.const.`**FORCE: bool**
If ignore `robots.txt` rules when crawling (c.f. `crawler()`).

Default `False`

Environ *DARC_FORCE*

`darc.const.`**CHECK: bool**
If check proxy and hostname before crawling (when calling `extract_links()`, `read_sitemap()` and `read_hosts()`).

If *CHECK_NG* is `True`, then this environment variable will be always set as `True`.

Default `False`

Environ *DARC_CHECK*

`darc.const.`**CHECK_NG: bool**
If check content type through `HEAD` requests before crawling (when calling `extract_links()`, `read_sitemap()` and `read_hosts()`).

Default `False`

Environ *DARC_CHECK_CONTENT_TYPE*

`darc.const.`**ROOT: str**
The root folder of the project.

`darc.const.`**CWD = '.'**
The current working direcory.

`darc.const.`**DARC_CPU: int**
Number of concurrent processes. If not provided, then the number of system CPUs will be used.

Default `None`

Environ *DARC_CPU*

`darc.const.`**FLAG_MP: bool**
If enable *multiprocessing* support.

Default `True`

Environ *DARC_MULTIPROCESSING*

`darc.const.`**FLAG_TH: bool**
If enable *multithreading* support.

Default `False`

Environ *DARC_MULTITHREADING*

Note: *FLAG_MP* and *FLAG_TH* can **NOT** be toggled at the same time.

---

darc.const.**DARC_USER: str**
Non-root user for proxies.

Default current login user (c.f. `getpass.getuser()`)

Environ *DARC_USER*

2.3.3 Data Storage

See also:

See `darc.db` for more information about database integration.

darc.const.**REDIS: redis.Redis**
URL to the Redis database.

Default `redis://127.0.0.1`

Environ *REDIS_URL*

darc.const.**DB: peewee.Database**
URL to the RDS storage.

Default `sqlite://{PATH_DB}/darc.db`

Environ :envvar`DB_URL`

darc.const.**DB: peewee.Database**
URL to the data submission storage.

Default `sqlite://{PATH_DB}/darcweb.db`

Environ :envvar`DB_URL`

darc.const.**FLAG_DB: bool**
Flag if uses RDS as the task queue backend. If *REDIS_URL* is provided, then `False`; else, `True`.

darc.const.**PATH_DB: str**
Path to data storage.

Default `data`

Environ *PATH_DATA*

See also:

See `darc.save` for more information about source saving.

darc.const.**PATH_MISC = '{PATH_DB}/misc/'**
Path to miscellaneous data storage, i.e. `misc` folder under the root of data storage.

See also:

- *darc.const.PATH_DB*

darc.const.**PATH_LN = '{PATH_DB}/link.csv'**
Path to the link CSV file, `link.csv`.

See also:

- *darc.const.PATH_DB*
- `darc.save.save_link`

darc.const.**PATH_ID = '{PATH_DB}/darc.pid'**
Path to the process ID file, `darc.pid`.

See also:

- *darc.const.PATH_DB*
- `darc.const.getpid()`

2.3.4 Web Crawlers

darc.const.**DARC_WAIT: Optional[float]**
Time interval between each round when the `requests` and/or `selenium` database are empty.

Default `60`

Environ *DARC_WAIT*

darc.const.**TIME_CACHE: float**
Time delta for caches in seconds.

The *darc* project supports *caching* for fetched files. *TIME_CACHE* will specify for how log the fetched files will be cached and **NOT** fetched again.

Note: If *TIME_CACHE* is `None` then caching will be marked as *forever*.

Default `60`

Environ *TIME_CACHE*

darc.const.**SE_WAIT: float**
Time to wait for `selenium` to finish loading pages.

Note: Internally, `selenium` will wait for the browser to finish loading the pages before return (i.e. the web API event `DOMContentLoaded`). However, some extra scripts may take more time running after the event.

Default `60`

Environ *SE_WAIT*

darc.const.**SE_EMPTY = '<html><head></head><body></body></html>'**
The empty page from `selenium`.

See also:

- `darc.crawl.loader()`

2.3.5 White / Black Lists

darc.const.**LINK_WHITE_LIST: List[re.Pattern]**
White list of hostnames should be crawled.

Default `[]`

Environ *LINK_WHITE_LIST*

Note: Regular expressions are supported.

darc.const.**LINK_BLACK_LIST: List[re.Pattern]**
Black list of hostnames should be crawled.

Default `[]`

Environ *LINK_BLACK_LIST*

Note: Regular expressions are supported.

darc.const.**LINK_FALLBACK: bool**
Fallback value for `match_host()`.

Default `False`

Environ *LINK_FALLBACK*

darc.const.**MIME_WHITE_LIST: List[re.Pattern]**
White list of content types should be crawled.

Default `[]`

Environ *MIME_WHITE_LIST*

Note: Regular expressions are supported.

darc.const.**MIME_BLACK_LIST: List[re.Pattern]**
Black list of content types should be crawled.

Default `[]`

Environ *MIME_BLACK_LIST*

Note: Regular expressions are supported.

darc.const.**MIME_FALLBACK: bool**
Fallback value for `match_mime()`.

Default `False`

Environ *MIME_FALLBACK*

darc.const.**PROXY_WHITE_LIST: List[str]**
White list of proxy types should be crawled.

Default `[]`

Environ *PROXY_WHITE_LIST*

Note: The proxy types are **case insensitive**.

darc.const.**PROXY_BLACK_LIST: List[str]**
Black list of proxy types should be crawled.

Default `[]`

Environ *PROXY_BLACK_LIST*

Note: The proxy types are **case insensitive**.

darc.const.**PROXY_FALLBACK: bool**
Fallback value for `match_proxy()`.

Default `False`

Environ *PROXY_FALLBACK*

2.4 Custom Exceptions

The *render_error()* function can be used to render multi-line error messages with `stem.util.term` colours.

The *darc* project provides following custom exceptions:

- *LinkNoReturn*
- *UnsupportedLink*
- *UnsupportedPlatform*
- *UnsupportedProxy*
- *WorkerBreak*

Note: All exceptions are inherited from *_BaseException*.

The *darc* project provides following custom warnings:

- *TorBootstrapFailed*
- *I2PBootstrapFailed*
- *ZeroNetBootstrapFailed*
- *FreenetBootstrapFailed*
- *APIRequestFailed*
- *SiteNotFoundWarning*
- *LockWarning*
- *TorRenewFailed*
- *RedisCommandFailed*
- *HookExecutionFailed*

Note: All warnings are inherited from *_BaseWarning*.

exception darc.error.**APIRequestFailed**
Bases: *darc.error._BaseWarning*

API submit failed.

exception darc.error.**DatabaseOperaionFailed**
Bases: *darc.error._BaseWarning*

Database operation execution failed.

exception darc.error.**FreenetBootstrapFailed**
Bases: *darc.error._BaseWarning*

Freenet bootstrap process failed.

exception darc.error.**HookExecutionFailed**
Bases: *darc.error._BaseWarning*

Failed to execute hook function.

exception darc.error.**I2PBootstrapFailed**
Bases: *darc.error._BaseWarning*

I2P bootstrap process failed.

exception darc.error.**LinkNoReturn**(*link=None*, *, *drop=True*)
Bases: *darc.error._BaseException*

The link has no return value from the hooks.

Parameters

- **link** (*darc.link.Link*) – Original link object.
- **drop** (*bool*) –

Keyword Arguments **drop** – If drops the `link` from task queues.

Return type None

__init__(*link=None*, *, *drop=True*)
Initialize self. See help(type(self)) for accurate signature.

Parameters **drop** (*bool*) –

Return type None

exception darc.error.**LockWarning**
Bases: *darc.error._BaseWarning*

Failed to acquire Redis lock.

exception darc.error.**RedisCommandFailed**
Bases: *darc.error._BaseWarning*

Redis command execution failed.

exception darc.error.**SiteNotFoundWarning**
Bases: *darc.error._BaseWarning*, `ImportWarning`

Site customisation not found.

exception darc.error.**TorBootstrapFailed**
Bases: *darc.error._BaseWarning*

Tor bootstrap process failed.

exception darc.error.**TorRenewFailed**
Bases: *darc.error._BaseWarning*

Tor renew request failed.

exception darc.error.**UnsupportedLink**
Bases: *darc.error._BaseException*

The link is not supported.

exception darc.error.**UnsupportedPlatform**
Bases: *darc.error._BaseException*

The platform is not supported.

exception darc.error.**UnsupportedProxy**
Bases: *darc.error._BaseException*

The proxy is not supported.

exception darc.error.**WorkerBreak**
Bases: *darc.error._BaseException*

Break from the worker loop.

exception darc.error.**ZeroNetBootstrapFailed**
Bases: *darc.error._BaseWarning*

ZeroNet bootstrap process failed.

exception darc.error.**_BaseException**
Bases: Exception

Base exception class for *darc* module.

exception darc.error.**_BaseWarning**
Bases: Warning

Base warning for *darc* module.

darc.error.**render_error**(*message*, *colour*)
Render error message.

The function wraps the stem.util.term.format() function to provide multi-line formatting support.

Parameters

- **message** (*AnyStr*) – Multi-line message to be rendered with colour.
- **colour** (*stem.util.term.Color*) – Front colour of text, c.f. stem.util.term.Color.

Returns The rendered error message.

Return type str

2.5 Data Models

The *darc.model* module contains all data models defined for the *darc* project, including RDS-based task queue and data submission.

2.5.1 Task Queues

The *darc.model.tasks* module defines the data models required for the task queue of *darc*.

See also:

Please refer to `darc.db` module for more information about the task queues.

2.5.2 Submission Data Models

The *darc.model.web* module defines the data models to store the data crawled from the *darc* project.

See also:

Please refer to `darc.submit` module for more information about data submission.

As the websites can be sometimes irritating for their anti-robots verification, login requirements, etc., the *darc* project also privides hooks to customise crawling behaviours around both `requests` and `selenium`.

See also:

Such customisation, as called in the *darc* project, site hooks, is site specific, user can set up your own hooks unto a certain site, c.f. *darc.sites* for more information.

Still, since the network is a world full of mysteries and miracles, the speed of crawling will much depend on the response speed of the target website. To boost up, as well as meet the system capacity, the *darc* project introduced multiprocessing, multithreading and the fallback slowest single-threaded solutions when crawling.

Note: When rendering the target website using `selenium` powered by the renown Google Chrome, it will require much memory to run. Thus, the three solutions mentioned above would only toggle the behaviour around the use of `selenium`.

To keep the *darc* project as it is a swiss army knife, only the main entrypoint function `darc.process.process()` is exported in global namespace (and renamed to `darc.darc()`), see below:

And we also exported the necessary hook registration functions to the global namespace, see below:

For more information on the hooks, please refer to the *customisation* documentations.

CHAPTER

THREE

CONFIGURATION

The *darc* project is generally configurable through numerous environment variables. Below is the full list of supported environment variables you may use to configure the behaviour of *darc*.

3.1 General Configurations

DARC_REBOOT

Type bool (int)

Default 0

If exit the program after first round, i.e. crawled all links from the requests link database and loaded all links from the selenium link database.

This can be useful especially when the capacity is limited and you wish to save some space before continuing next round. See *Docker integration* for more information.

DARC_DEBUG

Type bool (int)

Default 0

If run the program in debugging mode.

DARC_VERBOSE

Type bool (int)

Default 0

If run the program in verbose mode. If DARC_DEBUG is True, then the verbose mode will be always enabled.

DARC_FORCE

Type bool (int)

Default 0

If ignore robots.txt rules when crawling (c.f. crawler()).

DARC_CHECK

Type bool (int)

Default 0

If check proxy and hostname before crawling (when calling `extract_links()`, `read_sitemap()` and `read_hosts()`).

If `DARC_CHECK_CONTENT_TYPE` is `True`, then this environment variable will be always set as `True`.

DARC_CHECK_CONTENT_TYPE

Type `bool` (`int`)

Default `0`

If check content type through `HEAD` requests before crawling (when calling `extract_links()`, `read_sitemap()` and `read_hosts()`).

DARC_URL_PAT

Type List[Tuple[`str`, `int`]] (JSON)

Default `[]`

Regular expression patterns to match all reasonable URLs.

The environment variable should be **JSON** encoded, as an *array* of *two-element pairs*. In each pair, it contains one Python regular expression string (`str`) as described in the builtin `re` module and one numeric value (`int`) representing the flags as defined in the builtin `re` module as well.

Important: The patterns **must** have a named match group `url`, e.g. `(?P<url>bitcoin:\w+)` so that the function can extract matched URLs from the given pattern.

DARC_CPU

Type `int`

Default `None`

Number of concurrent processes. If not provided, then the number of system CPUs will be used.

DARC_MULTIPROCESSING

Type `bool` (`int`)

Default `1`

If enable *multiprocessing* support.

DARC_MULTITHREADING

Type `bool` (`int`)

Default `0`

If enable *multithreading* support.

Note: `DARC_MULTIPROCESSING` and `DARC_MULTITHREADING` can **NOT** be toggled at the same time.

DARC_USER

Type `str`

Default current login user (c.f. `getpass.getuser()`)

Non-root user for proxies.

3.2 Data Storage

See also:

See `darc.save` for more information about source saving.

See `darc.db` for more information about database integration.

PATH_DATA

Type `str` (path)

Default `data`

Path to data storage.

REDIS_URL

Type `str` (url)

Default `redis://127.0.0.1`

URL to the Redis database.

DB_URL

Type `str` (url)

URL to the RDS storage.

Important: The task queues will be saved to `darc` database; the data submittsion will be saved to `darcweb` database.

Thus, when providing this environment variable, please do **NOT** specify the database name.

DARC_BULK_SIZE

Type `int`

Default `100`

Bulk size for updating databases.

See also:

- `darc.db.save_requests()`
- `darc.db.save_selenium()`

LOCK_TIMEOUT

Type `float`

Default `10`

Lock blocking timeout.

Note: If is an infinit `inf`, no timeout will be applied.

See also:

Get a lock from `darc.db.get_lock()`.

DARC_MAX_POOL

Type int

Default 1_000

Maximum number of links loaded from the database.

Note: If is an infinit inf, no limit will be applied.

See also:

- darc.db.load_requests()
- darc.db.load_selenium()

REDIS_LOCK

Type bool (int)

Default 0

If use Redis (Lua) lock to ensure process/thread-safely operations.

See also:

Toggles the behaviour of darc.db.get_lock().

RETRY_INTERVAL

Type int

Default 10

Retry interval between each Redis command failure.

Note: If is an infinit inf, no interval will be applied.

See also:

Toggles the behaviour of darc.db.redis_command().

3.3 Web Crawlers

DARC_WAIT

Type float

Default 60

Time interval between each round when the requests and/or selenium database are empty.

DARC_SAVE

Type bool (int)

Default 0

If save processed link back to database.

Note: If *DARC_SAVE* is `True`, then *DARC_SAVE_REQUESTS* and *DARC_SAVE_SELENIUM* will be forced to be `True`.

See also:

See `darc.db` for more information about link database.

DARC_SAVE_REQUESTS

Type `bool` (`int`)

Default `0`

If save `crawler()` crawled link back to `requests` database.

See also:

See `darc.db` for more information about link database.

DARC_SAVE_SELENIUM

Type `bool` (`int`)

Default `0`

If save `loader()` crawled link back to `selenium` database.

See also:

See `darc.db` for more information about link database.

TIME_CACHE

Type `float`

Default `60`

Time delta for caches in seconds.

The *darc* project supports *caching* for fetched files. `TIME_CACHE` will specify for how log the fetched files will be cached and **NOT** fetched again.

Note: If `TIME_CACHE` is `None` then caching will be marked as *forever*.

SE_WAIT

Type `float`

Default `60`

Time to wait for `selenium` to finish loading pages.

Note: Internally, `selenium` will wait for the browser to finish loading the pages before return (i.e. the web API event `DOMContentLoaded`). However, some extra scripts may take more time running after the event.

CHROME_BINARY_LOCATION

Type `str`

Default `google-chrome`

Path to the Google Chrome binary location.

Note: This environment variable is mandatory for non *macOS* and/or *Linux* systems.

See also:

See `darc.selenium` for more information.

3.4 White / Black Lists

LINK_WHITE_LIST

Type `List[str]` (JSON)

Default `[]`

White list of hostnames should be crawled.

Note: Regular expressions are supported.

LINK_BLACK_LIST

Type `List[str]` (JSON)

Default `[]`

Black list of hostnames should be crawled.

Note: Regular expressions are supported.

LINK_FALLBACK

Type `bool` (`int`)

Default `0`

Fallback value for `match_host()`.

MIME_WHITE_LIST

Type `List[str]` (JSON)

Default `[]`

White list of content types should be crawled.

Note: Regular expressions are supported.

MIME_BLACK_LIST

Type `List[str]` (JSON)

Default `[]`

Black list of content types should be crawled.

Note: Regular expressions are supported.

MIME_FALLBACK

Type bool (int)

Default 0

Fallback value for match_mime().

PROXY_WHITE_LIST

Type List[str] (JSON)

Default []

White list of proxy types should be crawled.

Note: The proxy types are **case insensitive**.

PROXY_BLACK_LIST

Type List[str] (JSON)

Default []

Black list of proxy types should be crawled.

Note: The proxy types are **case insensitive**.

PROXY_FALLBACK

Type bool (int)

Default 0

Fallback value for match_proxy().

Note: If provided, LINK_WHITE_LIST, LINK_BLACK_LIST, MIME_WHITE_LIST, MIME_BLACK_LIST, PROXY_WHITE_LIST and PROXY_BLACK_LIST should all be JSON encoded strings.

3.5 Data Submission

SAVE_DB

Type bool

Default True

Save submitted data to database.

API_RETRY

Type int

Default 3

Retry times for API submission when failure.

API_NEW_HOST

Type str

Default None

API URL for submit_new_host().

API_REQUESTS

Type str

Default None

API URL for submit_requests().

API_SELENIUM

Type str

Default None

API URL for submit_selenium().

Note: If API_NEW_HOST, API_REQUESTS and API_SELENIUM is None, the corresponding submit function will save the JSON data in the path specified by PATH_DATA.

3.6 Tor Proxy Configuration

DARC_TOR

Type bool (int)

Default 1

If manage the Tor proxy through *darc*.

TOR_PORT

Type int

Default 9050

Port for Tor proxy connection.

TOR_CTRL

Type int

Default 9051

Port for Tor controller connection.

TOR_PASS

Type str

Default None

Tor controller authentication token.

Note: If not provided, it will be requested at runtime.

TOR_RETRY

Type int

Default 3

Retry times for Tor bootstrap when failure.

TOR_WAIT

Type float

Default 90

Time after which the attempt to start Tor is aborted.

Note: If not provided, there will be **NO** timeouts.

TOR_CFG

Type Dict[str, Any] (JSON)

Default {}

Tor bootstrap configuration for stem.process.launch_tor_with_config().

Note: If provided, it should be a JSON encoded string.

3.7 I2P Proxy Configuration

DARC_I2P

Type bool (int)

Default 1

If manage the I2P proxy through *darc*.

I2P_PORT

Type int

Default 4444

Port for I2P proxy connection.

I2P_RETRY

Type int

Default 3

Retry times for I2P bootstrap when failure.

I2P_WAIT

Type float

Default 90

Time after which the attempt to start I2P is aborted.

Note: If not provided, there will be **NO** timeouts.

I2P_ARGS

Type str (Shell)

Default ''

I2P bootstrap arguments for `i2prouter start`.

If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

Note: The command will be run as DARC_USER, if current user (c.f. `getpass.getuser()`) is *root*.

3.8 ZeroNet Proxy Configuration

DARC_ZERONET

Type bool (int)

Default 1

If manage the ZeroNet proxy through *darc*.

ZERONET_PORT

Type int

Default 4444

Port for ZeroNet proxy connection.

ZERONET_RETRY

Type int

Default 3

Retry times for ZeroNet bootstrap when failure.

ZERONET_WAIT

Type float

Default 90

Time after which the attempt to start ZeroNet is aborted.

Note: If not provided, there will be **NO** timeouts.

ZERONET_PATH

Type str (path)

Default `/usr/local/src/zeronet`

Path to the ZeroNet project.

ZERONET_ARGS

Type `str` (Shell)

Default `''`

ZeroNet bootstrap arguments for `ZeroNet.sh main`.

Note: If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

3.9 Freenet Proxy Configuration

DARC_FREENET

Type `bool` (`int`)

Default `1`

If manage the Freenet proxy through *darc*.

FREENET_PORT

Type `int`

Default `8888`

Port for Freenet proxy connection.

FREENET_RETRY

Type `int`

Default `3`

Retry times for Freenet bootstrap when failure.

FREENET_WAIT

Type `float`

Default `90`

Time after which the attempt to start Freenet is aborted.

Note: If not provided, there will be **NO** timeouts.

FREENET_PATH

Type `str` (path)

Default `/usr/local/src/freenet`

Path to the Freenet project.

FREENET_ARGS

Type `str` (Shell)

Default `''`

Freenet bootstrap arguments for `run.sh start`.

If provided, it should be parsed as command line arguments (c.f. `shlex.split()`).

Note: The command will be run as `DARC_USER`, if current user (c.f. `getpass.getuser()`) is *root*.

CHAPTER

FOUR

CUSTOMISATIONS

Currently, *darc* provides three major customisation points, besides the various *environment variables*.

4.1 Hooks between Rounds

See also:

See `darc.process.register()` for technical information.

As the workers are defined as indefinite loops, we introduced the *hooks between rounds* to be called at end of each loop. Such hook functions can process all links that had been crawled and/or loaded in the past round, or to indicate the end of the indefinite loop, so that we can stop the workers in an elegant way.

A typical hook function can be defined as following:

```python
from darc.error import WorkerBreak
from darc.process import register


def dummy_hook(node_type, link_pool):
    """A sample hook function that prints the processed links
    in the past round and informs the work to quit.

    Args:
        node_type (Literal['crawler', 'loader']): Type of worker node.
        link_pool (List[darc.link.Link]): List of processed links.

    Returns:
        NoReturn: The hook function will never return, though return
            values will be ignored anyway.

    Raises:
        darc.error.WorkerBreak: Inform the work to quit after this round.

    """
    if node_type == 'crawler':
        verb = 'crawled'
    elif node_type == 'loader':
        verb = 'loaded'
    else:
        raise ValueError('unknown type of worker node: %s' % node_type)

    for link in link_pool:
        print('We just %s the link: %s' % (verb, link.url))
```

(continues on next page)

(continued from previous page)

```
    raise WorkerBreak


# register the hook function
register(dummy_hook)
```

4.2 Custom Proxy

See also:

- *How to implement a custom proxy middleware?*
- See `darc.proxy.register()` for technical information.

Sometimes, we need proxies to connect to certain targers, such as the Tor network and I2P proxy. *darc* decides if it need to use a proxy for connection based on the `proxy` value of the target link.

By default, *darc* uses *no proxy* for `requests` sessions and `selenium` drivers. However, you may use your own proxies by registering and/or customising the corresponding factory functions.

A typical factory function pair (e.g., for Socks5 proxy) can be defined as following:

```
import requests
import requests_futures.sessions
import selenium.webdriver
import selenium.webdriver.common.proxy
from darc.const import DARC_CPU
from darc.proxy import register
from darc.requests import default_user_agent
from darc.selenium import BINARY_LOCATION


def socks5_session(futures=False):
    """Socks5 proxy session.

    Args:
        futures: If returns a :class:`requests_futures.FuturesSession`.

    Returns:
        Union[requests.Session, requests_futures.FuturesSession]:
        The session object with Socks5 proxy settings.

    """
    if futures:
        session = requests_futures.sessions.FuturesSession(max_workers=DARC_CPU)
    else:
        session = requests.Session()

    session.headers['User-Agent'] = default_user_agent(proxy='Socks5')
    session.proxies.update({
        'http': 'socks5h://localhost:9293',
        'https': 'socks5h://localhost:9293',
    })
    return session
```

(continues on next page)

(continued from previous page)

```
def socks5_driver():
    """Socks5 proxy driver.

    Returns:
        selenium.webdriver.Chrome: The web driver object with Socks5 proxy settings.

    """
    options = selenium.webdriver.ChromeOptions()
    options.binary_location = BINARY_LOCATION
    options.add_argument('--proxy-server=socks5://localhost:9293')
    options.add_argument('--host-resolver-rules="MAP * ~NOTFOUND , EXCLUDE localhost"
↪')

    proxy = selenium.webdriver.Proxy()
    proxy.proxyType = selenium.webdriver.common.proxy.ProxyType.MANUAL
    proxy.http_proxy = 'socks5://localhost:9293'
    proxy.ssl_proxy = 'socks5://localhost:9293'

    capabilities = selenium.webdriver.DesiredCapabilities.CHROME.copy()
    proxy.add_to_capabilities(capabilities)

    driver = selenium.webdriver.Chrome(options=options,
                                       desired_capabilities=capabilities)
    return driver


# register proxy
register('socks5', socks5_session, socks5_driver)
```

4.3 Sites Customisation

See also:

- *How to implement a sites customisation?*
- See `darc.sites.register()` for technical information.

Since websites may require authentication and/or anti-robot checks, we need to insert certain cookies, animate some user interactions to bypass such requirements. *darc* decides which customisation to use based on the hostname, i.e. `host` value of the target link.

By default, *darc* uses `darc.sites.default` as the *no op* for both `requests` sessions and `selenium` drivers. However, you may use your own sites customisation by registering and/or customising the corresponding classes, which inherited from `BaseSite`.

A typical sites customisation class (for better demonstration) can be defined as following:

```
import time

from darc.const import SE_WAIT
from darc.sites import BaseSite, register


class MySite(BaseSite):
    """This is a site customisation class for demonstration purpose.
```

(continues on next page)

(continued from previous page)

```
    You may implement a module as well should you prefer."""

    #: List[str]: Hostnames the sites customisation is designed for.
    hostname = ['mysite.com', 'www.mysite.com']

    @staticmethod
    def crawler(session, link):
        """Crawler hook for my site.

        Args:
            session (requests.Session): Session object with proxy settings.
            link (darc.link.Link): Link object to be crawled.

        Returns:
            requests.Response: The final response object with crawled data.

        """
        # inject cookies
        session.cookies.set('SessionID', 'fake-session-id-value')

        response = session.get(link.url, allow_redirects=True)
        return response

    @staticmethod
    def loader(driver, link):
        """Loader hook for my site.

        Args:
            driver (selenium.webdriver.Chrome): Web driver object with proxy settings.
            link (darc.link.Link): Link object to be loaded.

        Returns:
            selenium.webdriver.Chrome: The web driver object with loaded data.

        """
        # land on login page
        driver.get('https://%s/login' % link.host)

        # animate login attempt
        form = driver.find_element_by_id('login-form')
        form.find_element_by_id('username').send_keys('admin')
        form.find_element_by_id('password').send_keys('p@ssd')
        form.click()

        driver.get(link.url)

        # wait for page to finish loading
        if SE_WAIT is not None:
            time.sleep(SE_WAIT)

        return driver


# register sites
register(MySite)
```

Important: Please note that you may raise *darc.error.LinkNoReturn* in the `crawler` and/or `loader` methods to indicate that such link should be ignored and removed from the task queues, e.g. `darc.sites.data`.

CHAPTER

FIVE

DOCKER INTEGRATION

The *darc* project is integrated with Docker and Compose. Though published to Docker Hub, you can still build by yourself.

Important: The `debug` image contains miscellaneous documents, i.e. whole repository in it; and pre-installed some useful tools for debugging, such as IPython, etc.

The Docker image is based on Ubuntu Bionic (18.04 LTS), setting up all Python dependencies for the *darc* project, installing Google Chrome (version 79.0.3945.36) and corresponding ChromeDriver, as well as installing and configuring Tor, I2P, ZeroNet, FreeNet, NoIP proxies.

Note: NoIP is currently not fully integrated in the *darc* due to misunderstanding in the configuration process. Contributions are welcome.

When building the image, there is an *optional* argument for setting up a *non-root* user, c.f. environment variable `DARC_USER` and module constant *DARC_USER*. By default, the username is `darc`.

```
FROM ubuntu:focal

LABEL org.opencontainers.image.title="darc" \
      org.opencontainers.image.description="Darkweb Crawler Project" \
      org.opencontainers.image.url="https://darc.jarryshaw.me/" \
      org.opencontainers.image.source="https://github.com/JarryShaw/darc" \
      org.opencontainers.image.version="0.9.2" \
      org.opencontainers.image.licenses='BSD 3-Clause "New" or "Revised" License'

STOPSIGNAL SIGINT
HEALTHCHECK --interval=1h --timeout=1m \
    CMD wget https://httpbin.org/get -O /dev/null || exit 1

ARG DARC_USER="darc"
ENV LANG="C.UTF-8" \
    LC_ALL="C.UTF-8" \
    PYTHONIOENCODING="UTF-8" \
    DEBIAN_FRONTEND="teletype" \
    DARC_USER="${DARC_USER}"
    # DEBIAN_FRONTEND="noninteractive"

COPY extra/retry.sh /usr/local/bin/retry
COPY extra/install.py /usr/local/bin/pty-install
COPY vendor/jdk-11.0.9_linux-x64_bin.tar.gz /var/cache/oracle-jdk11-installer-local/
```

(continues on next page)

(continued from previous page)

```
RUN set -x \
 && retry apt-get update \
 && retry apt-get install --yes --no-install-recommends \
        apt-utils \
 && retry apt-get install --yes --no-install-recommends \
        gcc \
        g++ \
        libmagic1 \
        make \
        software-properties-common \
        tar \
        unzip \
        zlib1g-dev \
 && retry add-apt-repository ppa:deadsnakes/ppa --yes \
 && retry add-apt-repository ppa:linuxuprising/java --yes \
 && retry add-apt-repository ppa:i2p-maintainers/i2p --yes
RUN retry apt-get update \
 && retry apt-get install --yes --no-install-recommends \
        python3.9-dev \
        python3-pip \
        python3-setuptools \
        python3-wheel \
 && ln -sf /usr/bin/python3.9 /usr/local/bin/python3
RUN retry pty-install --stdin '6\n70' apt-get install --yes --no-install-recommends \
        tzdata \
 && retry pty-install --stdin 'yes' apt-get install --yes \
        oracle-java11-installer-local
RUN retry apt-get install --yes --no-install-recommends \
        sudo \
 && adduser --disabled-password --gecos '' ${DARC_USER} \
 && adduser ${DARC_USER} sudo \
 && echo '%sudo ALL=(ALL) NOPASSWD:ALL' >> /etc/sudoers

## Tor
RUN retry apt-get install --yes --no-install-recommends tor
COPY extra/torrc.focal /etc/tor/torrc

## I2P
RUN retry apt-get install --yes --no-install-recommends i2p
COPY extra/i2p.focal /etc/defaults/i2p

## ZeroNet
COPY vendor/ZeroNet-linux-dist-linux64.tar.gz /tmp
RUN set -x \
 && cd /tmp \
 && tar xvpfz ZeroNet-linux-dist-linux64.tar.gz \
 && mv ZeroNet-linux-dist-linux64 /usr/local/src/zeronet
COPY extra/zeronet.focal.conf /usr/local/src/zeronet/zeronet.conf

## FreeNet
USER darc
COPY vendor/new_installer_offline.jar /tmp
RUN set -x \
 && cd /tmp \
 && ( pty-install --stdin '/home/darc/freenet\n1' java -jar new_installer_offline.jar␣
↪|| true ) \
```

(continues on next page)

(continued from previous page)

```
 && sudo mv /home/darc/freenet /usr/local/src/freenet
USER root

## NoIP
COPY vendor/noip-duc-linux.tar.gz /tmp
RUN set -x \
 && cd /tmp \
 && tar xvpfz noip-duc-linux.tar.gz \
 && mv noip-2.1.9-1 /usr/local/src/noip \
 && cd /usr/local/src/noip \
 && make
 # && make install

# # set up timezone
# RUN echo 'Asia/Shanghai' > /etc/timezone \
#  && rm -f /etc/localtime \
#  && ln -snf /usr/share/zoneinfo/Asia/Shanghai /etc/localtime \
#  && dpkg-reconfigure -f noninteractive tzdata

COPY vendor/chromedriver_linux64.zip \
     vendor/google-chrome-stable_current_amd64.deb /tmp/
RUN set -x \
 ## ChromeDriver
 && unzip -d /usr/bin /tmp/chromedriver_linux64.zip \
 && which chromedriver \
 ## Google Chrome
 && ( dpkg --install /tmp/google-chrome-stable_current_amd64.deb || true ) \
 && retry apt-get install --fix-broken --yes --no-install-recommends \
 && dpkg --install /tmp/google-chrome-stable_current_amd64.deb \
 && which google-chrome

# Using pip:
COPY requirements.txt /tmp
RUN python3 -m pip install -r /tmp/requirements.txt --no-cache-dir

RUN set -x \
 && rm -rf \
        ## APT repository lists
        /var/lib/apt/lists/* \
        ## Python dependencies
        /tmp/requirements.txt \
        /tmp/pip \
        ## ChromeDriver
        /tmp/chromedriver_linux64.zip \
        ## Google Chrome
        /tmp/google-chrome-stable_current_amd64.deb \
        ## Vendors
        /tmp/new_installer_offline.jar \
        /tmp/noip-duc-linux.tar.gz \
        /tmp/ZeroNet-linux-dist-linux64.tar.gz \
 #&& apt-get remove --auto-remove --yes \
 #       software-properties-common \
 #       unzip \
 && apt-get autoremove -y \
 && apt-get autoclean \
 && apt-get clean
```

(continues on next page)

(continued from previous page)

```
ENTRYPOINT [ "python3", "-m", "darc" ]
#ENTRYPOINT [ "bash", "/app/run.sh" ]
CMD [ "--help" ]

WORKDIR /app
COPY darc/ /app/darc/
COPY LICENSE \
     MANIFEST.in \
     README.rst \
     extra/run.sh \
     setup.cfg \
     setup.py \
     test_darc.py /app/
RUN python3 -m pip install -e .
```

Note:

- `retry` is a shell script for retrying the commands until success

```
#!/usr/bin/env bash

while true; do
    >&2 echo "+ $@"
    $@ && break
    >&2 echo "exit: $?"
done
>&2 echo "exit: 0"
```

- `pty-install` is a Python script simulating user input for APT package installation with `DEBIAN_FRONTEND` set as `Teletype`.

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""Install packages requiring interactions."""

import argparse
import os
import subprocess  # nosec: B404
import sys
import tempfile
from typing import TYPE_CHECKING

if TYPE_CHECKING:
    from argparse import ArgumentParser


def get_parser() -> 'ArgumentParser':
    """Argument parser."""
    parser = argparse.ArgumentParser('install',
                                     description='pseudo-interactive package installer
↪')

    parser.add_argument('-i', '--stdin', help='content for input')
    parser.add_argument('command', nargs=argparse.REMAINDER, help='command to execute
↪')
```

(continues on next page)

(continued from previous page)

```
    return parser


def main() -> int:
    """Entrypoint."""
    parser = get_parser()
    args = parser.parse_args()
    text = args.stdin.encode().decode('unicode_escape')

    path = tempfile.mktemp(prefix='install-')  # nosec: B306
    with open(path, 'w') as file:
        file.write(text)

    with open(path, 'r') as file:
        proc = subprocess.run(args.command, stdin=file)  # pylint: disable=subprocess-
↪run-check # nosec: B603

    os.remove(path)
    return proc.returncode


if __name__ == "__main__":
    sys.exit(main())
```

As always, you can also use Docker Compose to manage the *darc* image. Environment variables can be set as described in the configuration section.

```
version: '3'

services:
  crawler:
    image: jsnbzh/darc:latest
    build: &build
      context: .
      args:
        # non-root user
        DARC_USER: "darc"
    container_name: crawler
    #entrypoint: [ "bash", "/app/run.sh" ]
    command: [ "--type", "crawler",
               "--file", "/app/text/tor.txt",
               "--file", "/app/text/tor2web.txt",
               "--file", "/app/text/i2p.txt",
               "--file", "/app/text/zeronet.txt",
               "--file", "/app/text/freenet.txt",
               "--file", "/app/text/clinic.txt" ]
    environment:
      ## [PYTHON] force the stdout and stderr streams to be unbuffered
      PYTHONUNBUFFERED: 1
      # reboot mode
      DARC_REBOOT: 0
      # debug mode
      DARC_DEBUG: 0
      # verbose mode
      DARC_VERBOSE: 1
```

(continues on next page)

(continued from previous page)

```
      # force mode (ignore robots.txt)
      DARC_FORCE: 1
      # check mode (check proxy and hostname before crawling)
      DARC_CHECK: 1
      # check mode (check content type before crawling)
      DARC_CHECK_CONTENT_TYPE: 0
      # save mode
      DARC_SAVE: 0
      # save mode (for requests)
      DAVE_SAVE_REQUESTS: 0
      # save mode (for selenium)
      DAVE_SAVE_SELENIUM: 0
      # processes
      DARC_CPU: 16
      # multiprocessing
      DARC_MULTIPROCESSING: 1
      # multithreading
      DARC_MULTITHREADING: 0
      # time lapse
      DARC_WAIT: 60
      # bulk size
      DARC_BULK_SIZE: 1000
      # data storage
      PATH_DATA: "data"
      # save data submitssion
      SAVE_DB: 0
      # Redis URL
      REDIS_URL: 'redis://
↪:UCf7y123aHgaYeGnvLRasALjFfDVHGCz6KiR5Z0WC0DL4ExvSGw5SkcOxBywc0qtZBHVrSVx2QMGewXNP6qVow@redis
↪'
      # database URL
      #DB_URL: 'mysql://
↪root:b8y9dpz3MJSQtwnZIW77ydASBOYfzA7HJfugv77wLrWQzrjCx5m3spoaiqRi4kU52syYy2jxJZR3U2kwPkEVTA@db
↪'
      # max pool
      DARC_MAX_POOL: 10
      # Tor proxy & control port
      TOR_PORT: 9050
      TOR_CTRL: 9051
      # Tor management method
      TOR_STEM: 1
      # Tor authentication
      TOR_PASS: "16:B9D36206B5374B3F609045F9609EE670F17047D88FF713EFB9157EA39F"
      # Tor bootstrap retry
      TOR_RETRY: 10
      # Tor bootstrap wait
      TOR_WAIT: 90
      # Tor bootstrap config
      TOR_CFG: "{}"
      # I2P port
      I2P_PORT: 4444
      # I2P bootstrap retry
      I2P_RETRY: 10
      # I2P bootstrap wait
      I2P_WAIT: 90
      # I2P bootstrap config
      I2P_ARGS: ""
```

(continues on next page)

(continued from previous page)

```
      # ZeroNet port
      ZERONET_PORT: 43110
      # ZeroNet bootstrap retry
      ZERONET_RETRY: 10
      # ZeroNet project path
      ZERONET_PATH: "/usr/local/src/zeronet"
      # ZeroNet bootstrap wait
      ZERONET_WAIT: 90
      # ZeroNet bootstrap config
      ZERONET_ARGS: ""
      # Freenet port
      FREENET_PORT: 8888
      # Freenet bootstrap retry
      FREENET_RETRY: 0
      # Freenet project path
      FREENET_PATH: "/usr/local/src/freenet"
      # Freenet bootstrap wait
      FREENET_WAIT: 90
      # Freenet bootstrap config
      FREENET_ARGS: ""
      # time delta for caches in seconds
      TIME_CACHE: 2_592_000  # 30 days
      # time to wait for selenium
      SE_WAIT: 5
      # extract link pattern
      LINK_WHITE_LIST: '[
        ".*?\\.onion",
        ".*?\\.i2p", "127\\.0\\.0\\.1:7657", "localhost:7657", "127\\.0\\.0\\.1:7658",
↪ "localhost:7658",
        "127\\.0\\.0\\.1:43110", "localhost:43110",
        "127\\.0\\.0\\.1:8888", "localhost:8888"
      ]'
      # link black list
      LINK_BLACK_LIST: '[ "(.*\\.)?facebookcorewwwi\\.onion", "(.*\\.)?
↪nytimes3xbfgragh\\.onion" ]'
      # link fallback flag
      LINK_FALLBACK: 1
      # content type white list
      MIME_WHITE_LIST: '[ "text/html", "application/xhtml+xml" ]'
      # content type black list
      MIME_BLACK_LIST: '[ "text/css", "application/javascript", "text/json" ]'
      # content type fallback flag
      MIME_FALLBACK: 0
      # proxy type white list
      PROXY_WHITE_LIST: '[ "tor", "i2p", "freenet", "zeronet", "tor2web" ]'
      # proxy type black list
      PROXY_BLACK_LIST: '[ "null", "data" ]'
      # proxy type fallback flag
      PROXY_FALLBACK: 0
      # API retry times
      API_RETRY: 10
      # API URLs
      #API_NEW_HOST: 'https://example.com/api/new_host'
      #API_REQUESTS: 'https://example.com/api/requests'
      #API_SELENIUM: 'https://example.com/api/selenium'
    restart: "always"
    networks: &networks
```

(continues on next page)

(continued from previous page)

```
      - darc
    volumes: &volumes
      - ./text:/app/text
      - ./extra:/app/extra
      - /data/darc:/app/data

  loader:
    image: jsnbzh/darc:latest
    build: *build
    container_name: loader
    #entrypoint: [ "bash", "/app/run.sh" ]
    command: [ "--type", "loader" ]
    environment:
      ## [PYTHON] force the stdout and stderr streams to be unbuffered
      PYTHONUNBUFFERED: 1
      # reboot mode
      DARC_REBOOT: 0
      # debug mode
      DARC_DEBUG: 0
      # verbose mode
      DARC_VERBOSE: 1
      # force mode (ignore robots.txt)
      DARC_FORCE: 1
      # check mode (check proxy and hostname before crawling)
      DARC_CHECK: 1
      # check mode (check content type before crawling)
      DARC_CHECK_CONTENT_TYPE: 0
      # save mode
      DARC_SAVE: 0
      # save mode (for requests)
      DAVE_SAVE_REQUESTS: 0
      # save mode (for selenium)
      DAVE_SAVE_SELENIUM: 0
      # processes
      DARC_CPU: 1
      # multiprocessing
      DARC_MULTIPROCESSING: 0
      # multithreading
      DARC_MULTITHREADING: 0
      # time lapse
      DARC_WAIT: 60
      # data storage
      PATH_DATA: "data"
      # Redis URL
      REDIS_URL: 'redis://
↪:UCf7y123aHgaYeGnvLRasALjFfDVHGCz6KiR5Z0WC0DL4ExvSGw5SkcOxBywc0qtZBHVrSVx2QMGewXNP6qVow@redis
↪'
      # database URL
      #DB_URL: 'mysql://
↪root:b8y9dpz3MJSQtwnZIW77ydASBOYfzA7HJfugv77wLrWQzrjCx5m3spoaiqRi4kU52syYy2jxJZR3U2kwPkEVTA@db
↪'
      # max pool
      DARC_MAX_POOL: 10
      # save data submitssion
      SAVE_DB: 0
      # Tor proxy & control port
      TOR_PORT: 9050
```

(continues on next page)

(continued from previous page)

```
      TOR_CTRL: 9051
      # Tor management method
      TOR_STEM: 1
      # Tor authentication
      TOR_PASS: "16:B9D36206B5374B3F609045F9609EE670F17047D88FF713EFB9157EA39F"
      # Tor bootstrap retry
      TOR_RETRY: 10
      # Tor bootstrap wait
      TOR_WAIT: 90
      # Tor bootstrap config
      TOR_CFG: "{}"
      # I2P port
      I2P_PORT: 4444
      # I2P bootstrap retry
      I2P_RETRY: 10
      # I2P bootstrap wait
      I2P_WAIT: 90
      # I2P bootstrap config
      I2P_ARGS: ""
      # ZeroNet port
      ZERONET_PORT: 43110
      # ZeroNet bootstrap retry
      ZERONET_RETRY: 10
      # ZeroNet project path
      ZERONET_PATH: "/usr/local/src/zeronet"
      # ZeroNet bootstrap wait
      ZERONET_WAIT: 90
      # ZeroNet bootstrap config
      ZERONET_ARGS: ""
      # Freenet port
      FREENET_PORT: 8888
      # Freenet bootstrap retry
      FREENET_RETRY: 0
      # Freenet project path
      FREENET_PATH: "/usr/local/src/freenet"
      # Freenet bootstrap wait
      FREENET_WAIT: 90
      # Freenet bootstrap config
      FREENET_ARGS: ""
      # time delta for caches in seconds
      TIME_CACHE: 2_592_000  # 30 days
      # time to wait for selenium
      SE_WAIT: 5
      # extract link pattern
      LINK_WHITE_LIST: '[
        ".*?\\.onion",
        ".*?\\.i2p", "127\\.0\\.0\\.1:7657", "localhost:7657", "127\\.0\\.0\\.1:7658",
↪ "localhost:7658",
        "127\\.0\\.0\\.1:43110", "localhost:43110",
        "127\\.0\\.0\\.1:8888", "localhost:8888"
      ]'
      # link black list
      LINK_BLACK_LIST: '[ "(.*\\.)?facebookcorewwwi\\.onion", "(.*\\.)?
↪nytimes3xbfgragh\\.onion" ]'
      # link fallback flag
      LINK_FALLBACK: 1
      # content type white list
```

(continues on next page)

(continued from previous page)

```
      MIME_WHITE_LIST: '[ "text/html", "application/xhtml+xml" ]'
      # content type black list
      MIME_BLACK_LIST: '[ "text/css", "application/javascript", "text/json" ]'
      # content type fallback flag
      MIME_FALLBACK: 0
      # proxy type white list
      PROXY_WHITE_LIST: '[ "tor", "i2p", "freenet", "zeronet", "tor2web" ]'
      # proxy type black list
      PROXY_BLACK_LIST: '[ "null", "data" ]'
      # proxy type fallback flag
      PROXY_FALLBACK: 0
      # API retry times
      API_RETRY: 10
      # API URLs
      #API_NEW_HOST: 'https://example.com/api/new_host'
      #API_REQUESTS: 'https://example.com/api/requests'
      #API_SELENIUM: 'https://example.com/api/selenium'
    restart: "always"
    networks: *networks
    volumes: *volumes

# network settings
networks:
  darc:
    driver: bridge
```

Note: Should you wish to run *darc* in reboot mode, i.e. set *DARC_REBOOT* and/or *REBOOT* as `True`, you may wish to change the entrypoint to

```
bash /app/run.sh
```

where `run.sh` is a shell script wraps around *darc* especially for reboot mode.

```
#!/usr/bin/env bash

set -e

# time lapse
WAIT=${DARC_WAIT=10}

# signal handlers
trap '[ -f ${PATH_DATA}/darc.pid ] && kill -2 $(cat ${PATH_DATA}/darc.pid)' SIGINT␣
↪SIGTERM SIGKILL

# initialise
echo "+ Starting application..."
python3 -m darc $@
sleep ${WAIT}

# mainloop
while true; do
    echo "+ Restarting application..."
    python3 -m darc
    sleep ${WAIT}
done
```

In such scenario, you can customise your `run.sh` to, for instance, archive then upload current data crawled by *darc* to somewhere else and save up some disk space.

CHAPTER

SIX

WEB BACKEND DEMO

This is a demo of API for communication between the *darc* crawlers (`darc.submit`) and web UI.

See also:

Please refer to *data schema* for more information about the submission data.

Assuming the web UI is developed using the `Flask` microframework.

```python
# -*- coding: utf-8 -*-

import sys
from typing import TYPE_CHECKING

import flask

if TYPE_CHECKING:
    from flask import Response

# Flask application
app = flask.Flask(__file__)


@app.route('/api/new_host', methods=['POST'])
def new_host() -> 'Response':
    """When a new host is discovered, the :mod:`darc` crawler will submit the
    host information. Such includes ``robots.txt`` (if exists) and
    ``sitemap.xml`` (if any).

    Data format::

        {
            // partial flag - true / false
            "$PARTIAL$": ...,
            // force flag - true / false
            "$FORCE$": ...,
            // metadata of URL
            "[metadata]": {
                // original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>
                "url": ...,
                // proxy type - null / tor / i2p / zeronet / freenet
                "proxy": ...,
                // hostname / netloc, c.f. ``urllib.parse.urlparse``
                "host": ...,
                // base folder, relative path (to data root path ``PATH_DATA``) in␣
↪containter - <proxy>/<scheme>/<host>
```

(continues on next page)

(continued from previous page)

```
                "base": ...,
                // sha256 of URL as name for saved files (timestamp is in ISO format)
                //   JSON log as this one - <base>/<name>_<timestamp>.json
                //   HTML from requests - <base>/<name>_<timestamp>_raw.html
                //   HTML from selenium - <base>/<name>_<timestamp>.html
                //   generic data files - <base>/<name>_<timestamp>.dat
                "name": ...
            },
            // requested timestamp in ISO format as in name of saved file
            "Timestamp": ...,
            // original URL
            "URL": ...,
            // robots.txt from the host (if not exists, then ``null``)
            "Robots": {
                // path of the file, relative path (to data root path ``PATH_DATA``)␣
↪in container
                //   - <proxy>/<scheme>/<host>/robots.txt
                "path": ...,
                // content of the file (**base64** encoded)
                "data": ...,
            },
            // sitemaps from the host (if none, then ``null``)
            "Sitemaps": [
                {
                    // path of the file, relative path (to data root path ``PATH_
↪DATA``) in container
                    //   - <proxy>/<scheme>/<host>/sitemap_<name>.xml
                    "path": ...,
                    // content of the file (**base64** encoded)
                    "data": ...,
                },
                ...
            ],
            // hosts.txt from the host (if proxy type is ``i2p``; if not exists, then␣
↪``null``)
            "Hosts": {
                // path of the file, relative path (to data root path ``PATH_DATA``)␣
↪in container
                //   - <proxy>/<scheme>/<host>/hosts.txt
                "path": ...,
                // content of the file (**base64** encoded)
                "data": ...,
            }
        }

    """
    # JSON data from the request
    data = flask.request.json  # pylint: disable=unused-variable

    # do whatever processing needed
    ...

    return flask.make_response()


@app.route('/api/requests', methods=['POST'])
```

(continues on next page)

(continued from previous page)

```
def from_requests() -> 'Response':
    """When crawling, we'll first fetch the URl using ``requests``, to check
    its availability and to save its HTTP headers information. Such information
    will be submitted to the web UI.

    Data format::

        {
            // metadata of URL
            "[metadata]": {
                // original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>
                "url": ...,
                // proxy type - null / tor / i2p / zeronet / freenet
                "proxy": ...,
                // hostname / netloc, c.f. ``urllib.parse.urlparse``
                "host": ...,
                // base folder, relative path (to data root path ``PATH_DATA``) in␣
↪containter - <proxy>/<scheme>/<host>
                "base": ...,
                // sha256 of URL as name for saved files (timestamp is in ISO format)
                //   JSON log as this one - <base>/<name>_<timestamp>.json
                //   HTML from requests - <base>/<name>_<timestamp>_raw.html
                //   HTML from selenium - <base>/<name>_<timestamp>.html
                //   generic data files - <base>/<name>_<timestamp>.dat
                "name": ...
            },
            // requested timestamp in ISO format as in name of saved file
            "Timestamp": ...,
            // original URL
            "URL": ...,
            // request method
            "Method": "GET",
            // response status code
            "Status-Code": ...,
            // response reason
            "Reason": ...,
            // response cookies (if any)
            "Cookies": {
                ...
            },
            // session cookies (if any)
            "Session": {
                ...
            },
            // request headers (if any)
            "Request": {
                ...
            },
            // response headers (if any)
            "Response": {
                ...
            },
            // content type
            "Content-Type": ...,
            // requested file (if not exists, then ``null``)
            "Document": {
```

(continues on next page)

(continued from previous page)

```
                // path of the file, relative path (to data root path ``PATH_DATA``)␣
↪in container
                //   - <proxy>/<scheme>/<host>/<name>_<timestamp>_raw.html
                // or if the document is of generic content type, i.e. not HTML
                //   - <proxy>/<scheme>/<host>/<name>_<timestamp>.dat
                "path": ...,
                // content of the file (**base64** encoded)
                "data": ...,
            },
            // redirection history (if any)
            "History": [
                // same records as the original response
                {"...": "..."}
            ]
        }

    """
    # JSON data from the request
    data = flask.request.json  # pylint: disable=unused-variable

    # do whatever processing needed
    ...

    return flask.make_response()


@app.route('/api/selenium', methods=['POST'])
def from_selenium() -> 'Response':
    """After crawling with ``requests``, we'll then render the URl using
    ``selenium`` with Google Chrome and its driver, to provide a fully rendered
    web page. Such information will be submitted to the web UI.

    Note:
        This information is optional, only provided if the content type from
        ``requests`` is HTML, status code < 400, and HTML data not empty.

    Data format::

        {
            // metadata of URL
            "[metadata]": {
                // original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>
                "url": ...,
                // proxy type - null / tor / i2p / zeronet / freenet
                "proxy": ...,
                // hostname / netloc, c.f. ``urllib.parse.urlparse``
                "host": ...,
                // base folder, relative path (to data root path ``PATH_DATA``) in␣
↪containter - <proxy>/<scheme>/<host>
                "base": ...,
                // sha256 of URL as name for saved files (timestamp is in ISO format)
                //   JSON log as this one - <base>/<name>_<timestamp>.json
                //   HTML from requests - <base>/<name>_<timestamp>_raw.html
                //   HTML from selenium - <base>/<name>_<timestamp>.html
                //   generic data files - <base>/<name>_<timestamp>.dat
                "name": ...
```

(continues on next page)

(continued from previous page)

```
            },
            // requested timestamp in ISO format as in name of saved file
            "Timestamp": ...,
            // original URL
            "URL": ...,
            // rendered HTML document (if not exists, then ``null``)
            "Document": {
                // path of the file, relative path (to data root path ``PATH_DATA``)␣
↪in container
                //   - <proxy>/<scheme>/<host>/<name>_<timestamp>.html
                "path": ...,
                // content of the file (**base64** encoded)
                "data": ...,
            },
            // web page screenshot (if not exists, then ``null``)
            "Screenshot": {
                // path of the file, relative path (to data root path ``PATH_DATA``)␣
↪in container
                //   - <proxy>/<scheme>/<host>/<name>_<timestamp>.png
                "path": ...,
                // content of the file (**base64** encoded)
                "data": ...,
            }
        }

    """
    # JSON data from the request
    data = flask.request.json  # pylint: disable=unused-variable

    # do whatever processing needed
    ...

    return flask.make_response()


if __name__ == "__main__":
    sys.exit(app.run())  # type: ignore[func-returns-value]
```

CHAPTER

SEVEN

DATA MODELS DEMO

This is a demo of data models for database storage of the submitted data from the *darc* crawlers.

Assuming the database is using peewee as ORM and MySQL as backend.

```python
# -*- coding: utf-8 -*-
# pylint: disable=ungrouped-imports

import os
from typing import TYPE_CHECKING

import peewee
import playhouse.shortcuts

if TYPE_CHECKING:
    from datetime import datetime
    from typing import Any, Dict

# database client
DB = playhouse.db_url.connect(os.getenv('DB_URL', 'mysql://127.0.0.1'))


def table_function(model_class: peewee.Model) -> str:
    """Generate table name dynamically.

    The function strips ``Model`` from the class name and
    calls :func:`peewee.make_snake_case` to generate a
    proper table name.

    Args:
        model_class: Data model class.

    Returns:
        Generated table name.

    """
    name = model_class.__name__  # type: str
    if name.endswith('Model'):
        name = name[:-5]  # strip ``Model`` suffix
    return peewee.make_snake_case(name)


class BaseMeta:
    """Basic metadata for data models."""
```

(continues on next page)

(continued from previous page)

```
    #: Reference database storage (c.f. :class:`~darc.const.DB`).
    database = DB

    #: Generate table name dynamically (c.f. :func:`~darc.model.table_function`).
    table_function = table_function


class BaseModel(peewee.Model):
    """Base model with standard patterns.

    Notes:
        The model will implicitly have a :class:`~peewee.AutoField`
        attribute named as :attr:`id`.

    """

    #: Basic metadata for data models.
    Meta = BaseMeta

    def to_dict(self, keep_id: bool = False) -> 'Dict[str, Any]':
        """Convert record to :obj:`dict`.

        Args:
            keep_id: If keep the ID auto field.

        Returns:
            The data converted through :func:`playhouse.shortcuts.model_to_dict`.

        """
        data = playhouse.shortcuts.model_to_dict(self)
        if keep_id:
            return data

        if 'id' in data:
            del data['id']
        return data


class HostnameModel(BaseModel):
    """Data model for a hostname record."""

    #: Hostname (c.f. :attr:`link.host <darc.link.Link.host>`).
    hostname: str = peewee.TextField()
    #: Proxy type (c.f. :attr:`link.proxy <darc.link.Link.proxy>`).
    proxy: str = peewee.CharField(max_length=8)

    #: Timestamp of first ``new_host`` submission.
    discovery: 'datetime' = peewee.DateTimeField()
    #: Timestamp of last related submission.
    last_seen: 'datetime' = peewee.DateTimeField()


class RobotsModel(BaseModel):
    """Data model for ``robots.txt`` data."""

    #: Hostname (c.f. :attr:`link.host <darc.link.Link.host>`).
    host: 'HostnameModel' = peewee.ForeignKeyField(HostnameModel, backref='robots')
```

(continues on next page)

(continued from previous page)

```
    #: Timestamp of the submission.
    timestamp: 'datetime' = peewee.DateTimeField()

    #: Document data as :obj:`bytes`.
    data: bytes = peewee.BlobField()
    #: Path to the document.
    path: str = peewee.CharField()


class SitemapModel(BaseModel):
    """Data model for ``sitemap.xml`` data."""

    #: Hostname (c.f. :attr:`link.host <darc.link.Link.host>`).
    host: 'HostnameModel' = peewee.ForeignKeyField(HostnameModel, backref='sitemaps')
    #: Timestamp of the submission.
    timestamp: 'datetime' = peewee.DateTimeField()

    #: Document data as :obj:`bytes`.
    data: bytes = peewee.BlobField()
    #: Path to the document.
    path: str = peewee.CharField()


class HostsModel(BaseModel):
    """Data model for ``hosts.txt`` data."""

    #: Hostname (c.f. :attr:`link.host <darc.link.Link.host>`).
    host: 'HostnameModel' = peewee.ForeignKeyField(HostnameModel, backref='hosts')
    #: Timestamp of the submission.
    timestamp: 'datetime' = peewee.DateTimeField()

    #: Document data as :obj:`bytes`.
    data: bytes = peewee.BlobField()
    #: Path to the document.
    path: str = peewee.CharField()


class URLModel(BaseModel):
    """Data model for a requested URL."""

    #: Timestamp of last related submission.
    last_seen: 'datetime' = peewee.DateTimeField()
    #: Original URL (c.f. :attr:`link.url <darc.link.Link.url>`).
    url: str = peewee.TextField()

    #: Hostname (c.f. :attr:`link.host <darc.link.Link.host>`).
    host: HostnameModel = peewee.ForeignKeyField(HostnameModel, backref='urls')
    #: Proxy type (c.f. :attr:`link.proxy <darc.link.Link.proxy>`).
    proxy: str = peewee.CharField(max_length=8)

    #: Base path (c.f. :attr:`link.base <darc.link.Link.base>`).
    base: str = peewee.CharField()
    #: Link hash (c.f. :attr:`link.name <darc.link.Link.name>`).
    name: str = peewee.FixedCharField(max_length=64)


class RequestsDocumentModel(BaseModel):
```

(continues on next page)

(continued from previous page)

```
    """Data model for documents from ``requests`` submission."""

    #: Original URL (c.f. :attr:`link.url <darc.link.Link.url>`).
    url: 'URLModel' = peewee.ForeignKeyField(URLModel, backref='requests')

    #: Document data as :obj:`bytes`.
    data: bytes = peewee.BlobField()
    #: Path to the document.
    path: str = peewee.CharField()


class SeleniumDocumentModel(BaseModel):
    """Data model for documents from ``selenium`` submission."""

    #: Original URL (c.f. :attr:`link.url <darc.link.Link.url>`).
    url: 'URLModel' = peewee.ForeignKeyField(URLModel, backref='selenium')

    #: Document data as :obj:`bytes`.
    data: bytes = peewee.BlobField()
    #: Path to the document.
    path: str = peewee.CharField()
```

CHAPTER

EIGHT

SUBMISSION DATA SCHEMA

To better describe the submitted data, *darc* provides several JSON schema generated from `pydantic` models.

8.1 New Host Submission

The data submission from `darc.submit.submit_new_host()`.


```
{
  "title": "new_host",
  "description": "Data submission from :func:`darc.submit.submit_new_host`.",
  "type": "object",
  "properties": {
    "$PARTIAL$": {
      "title": "$Partial$",
      "description": "partial flag - true / false",
      "type": "boolean"
    },
    "$RELOAD$": {
      "title": "$Reload$",
      "description": "reload flag - true / false",
      "type": "boolean"
    },
    "[metadata]": {
      "title": "[Metadata]",
      "description": "metadata of URL",
      "allOf": [
        {
          "$ref": "#/definitions/Metadata"
        }
      ]
    },
    "Timestamp": {
      "title": "Timestamp",
      "description": "requested timestamp in ISO format as in name of saved file",
      "type": "string",
      "format": "date-time"
    },
    "URL": {
      "title": "Url",
      "description": "original URL",
      "minLength": 1,
      "maxLength": 65536,
      "format": "uri",
```


(continues on next page)

(continued from previous page)

```
      "type": "string"
    },
    "Robots": {
      "title": "Robots",
      "description": "robots.txt from the host (if not exists, then ``null``)",
      "allOf": [
        {
          "$ref": "#/definitions/RobotsDocument"
        }
      ]
    },
    "Sitemaps": {
      "title": "Sitemaps",
      "description": "sitemaps from the host (if none, then ``null``)",
      "type": "array",
      "items": {
        "$ref": "#/definitions/SitemapDocument"
      }
    },
    "Hosts": {
      "title": "Hosts",
      "description": "hosts.txt from the host (if proxy type is ``i2p``; if not␣
↪exists, then ``null``)",
      "allOf": [
        {
          "$ref": "#/definitions/HostsDocument"
        }
      ]
    }
  },
  "required": [
    "$PARTIAL$",
    "$RELOAD$",
    "[metadata]",
    "Timestamp",
    "URL"
  ],
  "definitions": {
    "Proxy": {
      "title": "Proxy",
      "description": "Proxy type.",
      "enum": [
        "null",
        "tor",
        "i2p",
        "zeronet",
        "freenet"
      ],
      "type": "string"
    },
    "Metadata": {
      "title": "metadata",
      "description": "Metadata of URL.",
      "type": "object",
      "properties": {
        "url": {
          "title": "Url",
```

(continues on next page)

(continued from previous page)

```
          "description": "original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>",
          "minLength": 1,
          "maxLength": 65536,
          "format": "uri",
          "type": "string"
        },
        "proxy": {
          "$ref": "#/definitions/Proxy"
        },
        "host": {
          "title": "Host",
          "description": "hostname / netloc, c.f. ``urllib.parse.urlparse``",
          "type": "string"
        },
        "base": {
          "title": "Base",
          "description": "base folder, relative path (to data root path ``PATH_
↪DATA``) in containter - <proxy>/<scheme>/<host>",
          "type": "string"
        },
        "name": {
          "title": "Name",
          "description": "sha256 of URL as name for saved files (timestamp is in ISO␣
↪format) - JSON log as this one: <base>/<name>_<timestamp>.json; - HTML from␣
↪requests: <base>/<name>_<timestamp>_raw.html; - HTML from selenium: <base>/<name>_
↪<timestamp>.html; - generic data files: <base>/<name>_<timestamp>.dat",
          "type": "string"
        }
      },
      "required": [
        "url",
        "proxy",
        "host",
        "base",
        "name"
      ]
    },
    "RobotsDocument": {
      "title": "RobotsDocument",
      "description": "``robots.txt`` document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/robots.txt",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
```

(continues on next page)

(continued from previous page)

```
        "data"
      ]
    },
    "SitemapDocument": {
      "title": "SitemapDocument",
      "description": "Sitemaps document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/sitemap_<name>.xml",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
        "data"
      ]
    },
    "HostsDocument": {
      "title": "HostsDocument",
      "description": "``hosts.txt`` document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/hosts.txt",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
        "data"
      ]
    }
  }
}
```

8.2 Requests Submission

The data submission from `darc.submit.submit_requests()`.


```
{
  "title": "requests",
  "description": "Data submission from :func:`darc.submit.submit_requests`.",
  "type": "object",
  "properties": {
    "$PARTIAL$": {
      "title": "$Partial$",
      "description": "partial flag - true / false",
      "type": "boolean"
    },
    "[metadata]": {
      "title": "[Metadata]",
      "description": "metadata of URL",
      "allOf": [
        {
          "$ref": "#/definitions/Metadata"
        }
      ]
    },
    "Timestamp": {
      "title": "Timestamp",
      "description": "requested timestamp in ISO format as in name of saved file",
      "type": "string",
      "format": "date-time"
    },
    "URL": {
      "title": "Url",
      "description": "original URL",
      "minLength": 1,
      "maxLength": 65536,
      "format": "uri",
      "type": "string"
    },
    "Method": {
      "title": "Method",
      "description": "request method",
      "type": "string"
    },
    "Status-Code": {
      "title": "Status-Code",
      "description": "response status code",
      "exclusiveMinimum": 0,
      "type": "integer"
    },
    "Reason": {
      "title": "Reason",
      "description": "response reason",
      "type": "string"
    },
    "Cookies": {
      "title": "Cookies",
      "description": "response cookies (if any)",
      "type": "object",
```


(continues on next page)

(continued from previous page)

```
      "additionalProperties": {
        "type": "string"
      }
    },
    "Session": {
      "title": "Session",
      "description": "session cookies (if any)",
      "type": "object",
      "additionalProperties": {
        "type": "string"
      }
    },
    "Request": {
      "title": "Request",
      "description": "request headers (if any)",
      "type": "object",
      "additionalProperties": {
        "type": "string"
      }
    },
    "Response": {
      "title": "Response",
      "description": "response headers (if any)",
      "type": "object",
      "additionalProperties": {
        "type": "string"
      }
    },
    "Content-Type": {
      "title": "Content-Type",
      "description": "content type",
      "pattern": "[a-zA-Z0-9.-]+/[a-zA-Z0-9.-]+",
      "type": "string"
    },
    "Document": {
      "title": "Document",
      "description": "requested file (if not exists, then ``null``)",
      "allOf": [
        {
          "$ref": "#/definitions/RequestsDocument"
        }
      ]
    },
    "History": {
      "title": "History",
      "description": "redirection history (if any)",
      "type": "array",
      "items": {
        "$ref": "#/definitions/HistoryModel"
      }
    }
  },
  "required": [
    "$PARTIAL$",
    "[metadata]",
    "Timestamp",
    "URL",
```

(continues on next page)

(continued from previous page)


```
    "Method",
    "Status-Code",
    "Reason",
    "Cookies",
    "Session",
    "Request",
    "Response",
    "Content-Type",
    "History"
  ],
  "definitions": {
    "Proxy": {
      "title": "Proxy",
      "description": "Proxy type.",
      "enum": [
        "null",
        "tor",
        "i2p",
        "zeronet",
        "freenet"
      ],
      "type": "string"
    },
    "Metadata": {
      "title": "metadata",
      "description": "Metadata of URL.",
      "type": "object",
      "properties": {
        "url": {
          "title": "Url",
          "description": "original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>",
          "minLength": 1,
          "maxLength": 65536,
          "format": "uri",
          "type": "string"
        },
        "proxy": {
          "$ref": "#/definitions/Proxy"
        },
        "host": {
          "title": "Host",
          "description": "hostname / netloc, c.f. ``urllib.parse.urlparse``",
          "type": "string"
        },
        "base": {
          "title": "Base",
          "description": "base folder, relative path (to data root path ``PATH_
↪DATA``) in containter - <proxy>/<scheme>/<host>",
          "type": "string"
        },
        "name": {
          "title": "Name",
          "description": "sha256 of URL as name for saved files (timestamp is in ISO␣
↪format) - JSON log as this one: <base>/<name>_<timestamp>.json; - HTML from␣
↪requests: <base>/<name>_<timestamp>_raw.html; - HTML from selenium: <base>/<name>_
↪<timestamp>.html; - generic data files: <base>/<name>_<timestamp>.dat",
```


(continues on next page)

(continued from previous page)


```
          "type": "string"
        }
      },
      "required": [
        "url",
        "proxy",
        "host",
        "base",
        "name"
      ]
    },
    "RequestsDocument": {
      "title": "RequestsDocument",
      "description": ":mod:`requests` document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/<name>_<timestamp>_raw.html; or if␣
↪the document is of generic content type, i.e. not HTML - <proxy>/<scheme>/<host>/
↪<name>_<timestamp>.dat",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
        "data"
      ]
    },
    "HistoryModel": {
      "title": "HistoryModel",
      "description": ":mod:`requests` history data.",
      "type": "object",
      "properties": {
        "URL": {
          "title": "Url",
          "description": "original URL",
          "minLength": 1,
          "maxLength": 65536,
          "format": "uri",
          "type": "string"
        },
        "Method": {
          "title": "Method",
          "description": "request method",
          "type": "string"
        },
        "Status-Code": {
          "title": "Status-Code",
          "description": "response status code",
          "exclusiveMinimum": 0,
```


(continues on next page)

(continued from previous page)

```
        "type": "integer"
      },
      "Reason": {
        "title": "Reason",
        "description": "response reason",
        "type": "string"
      },
      "Cookies": {
        "title": "Cookies",
        "description": "response cookies (if any)",
        "type": "object",
        "additionalProperties": {
          "type": "string"
        }
      },
      "Session": {
        "title": "Session",
        "description": "session cookies (if any)",
        "type": "object",
        "additionalProperties": {
          "type": "string"
        }
      },
      "Request": {
        "title": "Request",
        "description": "request headers (if any)",
        "type": "object",
        "additionalProperties": {
          "type": "string"
        }
      },
      "Response": {
        "title": "Response",
        "description": "response headers (if any)",
        "type": "object",
        "additionalProperties": {
          "type": "string"
        }
      },
      "Document": {
        "title": "Document",
        "description": "content of the file (**base64** encoded)",
        "type": "string"
      }
    },
    "required": [
      "URL",
      "Method",
      "Status-Code",
      "Reason",
      "Cookies",
      "Session",
      "Request",
      "Response",
      "Document"
    ]
  }
```

(continues on next page)

(continued from previous page)

```
  }
}
```

8.3 Selenium Submission

The data submission from `darc.submit.submit_selenium()`.

```
{
  "title": "selenium",
  "description": "Data submission from :func:`darc.submit.submit_requests`.",
  "type": "object",
  "properties": {
    "$PARTIAL$": {
      "title": "$Partial$",
      "description": "partial flag - true / false",
      "type": "boolean"
    },
    "[metadata]": {
      "title": "[Metadata]",
      "description": "metadata of URL",
      "allOf": [
        {
          "$ref": "#/definitions/Metadata"
        }
      ]
    },
    "Timestamp": {
      "title": "Timestamp",
      "description": "requested timestamp in ISO format as in name of saved file",
      "type": "string",
      "format": "date-time"
    },
    "URL": {
      "title": "Url",
      "description": "original URL",
      "minLength": 1,
      "maxLength": 65536,
      "format": "uri",
      "type": "string"
    },
    "Document": {
      "title": "Document",
      "description": "rendered HTML document (if not exists, then ``null``)",
      "allOf": [
        {
          "$ref": "#/definitions/SeleniumDocument"
        }
      ]
    },
    "Screenshot": {
      "title": "Screenshot",
      "description": "web page screenshot (if not exists, then ``null``)",
      "allOf": [
        {
```

(continues on next page)

(continued from previous page)

```
          "$ref": "#/definitions/ScreenshotDocument"
        }
      ]
    }
  },
  "required": [
    "$PARTIAL$",
    "[metadata]",
    "Timestamp",
    "URL"
  ],
  "definitions": {
    "Proxy": {
      "title": "Proxy",
      "description": "Proxy type.",
      "enum": [
        "null",
        "tor",
        "i2p",
        "zeronet",
        "freenet"
      ],
      "type": "string"
    },
    "Metadata": {
      "title": "metadata",
      "description": "Metadata of URL.",
      "type": "object",
      "properties": {
        "url": {
          "title": "Url",
          "description": "original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>",
          "minLength": 1,
          "maxLength": 65536,
          "format": "uri",
          "type": "string"
        },
        "proxy": {
          "$ref": "#/definitions/Proxy"
        },
        "host": {
          "title": "Host",
          "description": "hostname / netloc, c.f. ``urllib.parse.urlparse``",
          "type": "string"
        },
        "base": {
          "title": "Base",
          "description": "base folder, relative path (to data root path ``PATH_
↪DATA``) in containter - <proxy>/<scheme>/<host>",
          "type": "string"
        },
        "name": {
          "title": "Name",
          "description": "sha256 of URL as name for saved files (timestamp is in ISO␣
↪format) - JSON log as this one: <base>/<name>_<timestamp>.json; - HTML from␣
↪requests: <base>/<name>_<timestamp>_raw.html; - HTML from selenium: <base>/<name>_
↪<timestamp>.html; - generic data files: <base>/<name>_<timestamp>.dat",
```

(continues on next page)

(continued from previous page)

```
          "type": "string"
        }
      },
      "required": [
        "url",
        "proxy",
        "host",
        "base",
        "name"
      ]
    },
    "SeleniumDocument": {
      "title": "SeleniumDocument",
      "description": ":mod:`selenium` document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/<name>_<timestamp>.html",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
        "data"
      ]
    },
    "ScreenshotDocument": {
      "title": "ScreenshotDocument",
      "description": "Screenshot document data.",
      "type": "object",
      "properties": {
        "path": {
          "title": "Path",
          "description": "path of the file, relative path (to data root path ``PATH_
↪DATA``) in container - <proxy>/<scheme>/<host>/<name>_<timestamp>.png",
          "type": "string"
        },
        "data": {
          "title": "Data",
          "description": "content of the file (**base64** encoded)",
          "type": "string"
        }
      },
      "required": [
        "path",
        "data"
      ]
    }
  }
}
```

8.4 Model Definitions

```python
# -*- coding: utf-8 -*-
# pylint: disable=ungrouped-imports
"""JSON schema generator."""

from typing import TYPE_CHECKING

import pydantic.schema
from pydantic import BaseModel, Field

__all__ = ['NewHostModel', 'RequestsModel', 'SeleniumModel']

if TYPE_CHECKING:
    from datetime import datetime
    from enum import Enum
    from typing import Any, Dict, List, Optional

    from pydantic import AnyUrl, PositiveInt

    CookiesType = List[Dict[str, Any]]
    HeadersType = Dict[str, str]

    class Proxy(str, Enum):
        """Proxy type."""

        null = 'null'
        tor = 'tor'
        i2p = 'i2p'
        zeronet = 'zeronet'
        freenet = 'freenet'

################################################################################
# Miscellaneous auxiliaries
################################################################################


class Metadata(BaseModel):
    """Metadata of URL."""

    url: 'AnyUrl' = Field(
        description='original URL - <scheme>://<netloc>/<path>;<params>?<query>#
↪<fragment>')
    proxy: 'Proxy' = Field(
        description='proxy type - null / tor / i2p / zeronet / freenet')
    host: str = Field(
        description='hostname / netloc, c.f. ``urllib.parse.urlparse``')
    base: str = Field(
        description=('base folder, relative path (to data root path ``PATH_DATA``) in␣
↪containter '
                     '- <proxy>/<scheme>/<host>'))
    name: str = Field(
        description=('sha256 of URL as name for saved files (timestamp is in ISO␣
↪format) '
                     '- JSON log as this one: <base>/<name>_<timestamp>.json; '
                     '- HTML from requests: <base>/<name>_<timestamp>_raw.html; '
                     '- HTML from selenium: <base>/<name>_<timestamp>.html; '
```

(continues on next page)

(continued from previous page)

```
                    '- generic data files: <base>/<name>_<timestamp>.dat'))

    class Config:
        title = 'metadata'


class RobotsDocument(BaseModel):
    """``robots.txt`` document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/robots.txt'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


class SitemapDocument(BaseModel):
    """Sitemaps document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/sitemap_<name>.xml'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


class HostsDocument(BaseModel):
    """``hosts.txt`` document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/hosts.txt'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


class RequestsDocument(BaseModel):
    """:mod:`requests` document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/<name>_<timestamp>_raw.html; '
                     'or if the document is of generic content type, i.e. not HTML '
                     '- <proxy>/<scheme>/<host>/<name>_<timestamp>.dat'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


class HistoryModel(BaseModel):
    """:mod:`requests` history data."""

    URL: 'AnyUrl' = Field(
        description='original URL')
```

(continues on next page)

(continued from previous page)

```
    Method: str = Field(
        description='request method')
    status_code: 'PositiveInt' = Field(
        alias='Status-Code',
        description='response status code')
    Reason: str = Field(
        description='response reason')

    Cookies: 'CookiesType' = Field(
        description='response cookies (if any)')
    Session: 'CookiesType' = Field(
        description='session cookies (if any)')

    Request: 'HeadersType' = Field(
        description='request headers (if any)')
    Response: 'HeadersType' = Field(
        description='response headers (if any)')

    Document: str = Field(
        description='content of the file (**base64** encoded)')


class SeleniumDocument(BaseModel):
    """:mod:`selenium` document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/<name>_<timestamp>.html'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


class ScreenshotDocument(BaseModel):
    """Screenshot document data."""

    path: str = Field(
        description=('path of the file, relative path (to data root path ``PATH_
↪DATA``) in container '
                     '- <proxy>/<scheme>/<host>/<name>_<timestamp>.png'))
    data: str = Field(
        description='content of the file (**base64** encoded)')


################################################################################
# JSON schema definitions
################################################################################


class NewHostModel(BaseModel):
    """Data submission from :func:`darc.submit.submit_new_host`."""

    partial: bool = Field(
        alias='$PARTIAL$',
        description='partial flag - true / false')
    reload: bool = Field(
```

(continues on next page)

(continued from previous page)

```
        alias='$RELOAD$',
        description='reload flag - true / false')
    metadata: 'Metadata' = Field(
        alias='[metadata]',
        description='metadata of URL')

    Timestamp: 'datetime' = Field(
        description='requested timestamp in ISO format as in name of saved file')
    URL: 'AnyUrl' = Field(
        description='original URL')

    Robots: 'Optional[RobotsDocument]' = Field(
        description='robots.txt from the host (if not exists, then ``null``)')
    Sitemaps: 'Optional[List[SitemapDocument]]' = Field(
        description='sitemaps from the host (if none, then ``null``)')
    Hosts: 'Optional[HostsDocument]' = Field(
        description='hosts.txt from the host (if proxy type is ``i2p``; if not exists,
↪ then ``null``)')


    class Config:
        title = 'new_host'


class RequestsModel(BaseModel):
    """Data submission from :func:`darc.submit.submit_requests`."""

    partial: bool = Field(
        alias='$PARTIAL$',
        description='partial flag - true / false')
    metadata: 'Metadata' = Field(
        alias='[metadata]',
        description='metadata of URL')

    Timestamp: 'datetime' = Field(
        description='requested timestamp in ISO format as in name of saved file')
    URL: 'AnyUrl' = Field(
        description='original URL')

    Method: str = Field(
        description='request method')
    status_code: 'PositiveInt' = Field(
        alias='Status-Code',
        description='response status code')
    Reason: str = Field(
        description='response reason')

    Cookies: 'CookiesType' = Field(
        description='response cookies (if any)')
    Session: 'CookiesType' = Field(
        description='session cookies (if any)')

    Request: 'HeadersType' = Field(
        description='request headers (if any)')
    Response: 'HeadersType' = Field(
        description='response headers (if any)')
    content_type: str = Field(
```

(continues on next page)

(continued from previous page)

```
        alias='Content-Type',
        regex='[a-zA-Z0-9.-]+/[a-zA-Z0-9.-]+',
        description='content type')

    Document: 'Optional[RequestsDocument]' = Field(
        description='requested file (if not exists, then ``null``)')
    History: 'List[HistoryModel]' = Field(
        description='redirection history (if any)')

    class Config:
        title = 'requests'


class SeleniumModel(BaseModel):
    """Data submission from :func:`darc.submit.submit_requests`."""

    partial: bool = Field(
        alias='$PARTIAL$',
        description='partial flag - true / false')
    metadata: 'Metadata' = Field(
        alias='[metadata]',
        description='metadata of URL')

    Timestamp: 'datetime' = Field(
        description='requested timestamp in ISO format as in name of saved file')
    URL: 'AnyUrl' = Field(
        description='original URL')

    Document: 'Optional[SeleniumDocument]' = Field(
        description='rendered HTML document (if not exists, then ``null``)')
    Screenshot: 'Optional[ScreenshotDocument]' = Field(
        description='web page screenshot (if not exists, then ``null``)')

    class Config:
        title = 'selenium'


if __name__ == "__main__":
    import json
    import os

    os.makedirs('schema', exist_ok=True)

    with open('schema/new_host.schema.json', 'w') as file:
        print(NewHostModel.schema_json(indent=2), file=file)
    with open('schema/requests.schema.json', 'w') as file:
        print(RequestsModel.schema_json(indent=2), file=file)
    with open('schema/selenium.schema.json', 'w') as file:
        print(SeleniumModel.schema_json(indent=2), file=file)

    schema = pydantic.schema.schema([NewHostModel, RequestsModel, SeleniumModel],
                                    title='DARC Data Submission JSON Schema')
    with open('schema/darc.schema.json', 'w') as file:
        json.dump(schema, file, indent=2)
```

CHAPTER

NINE

AUXILIARY SCRIPTS

Since the *darc* project can be deployed through *Docker Integration*, we provided some auxiliary scripts to help with the deployment.

9.1 Health Check

File location

- Entry point: `extra/healthcheck.py`
- System V service: `extra/healthcheck.service`

```
usage: healthcheck [-h] [-f FILE] [-i INTERVAL] ...

health check running container

positional arguments:
  services              name of services

optional arguments:
  -h, --help            show this help message and exit
  -f FILE, --file FILE  path to compose file
  -i INTERVAL, --interval INTERVAL
                        interval (in seconds) of health check
```

This script will watch the running status of containers managed by Docker Compose. If the containers are stopped or of *unhealthy* status, it will bring the containers back alive.

Also, as the internal program may halt unexpectedly whilst the container remains *healthy*, the script will watch if the program is still active through its output messages. If inactive, the script will restart the containers.

9.2 Upload API Submission Files

File location

- Entry point: `extra/upload.py`
- Helper script: `extra/upload.sh`
- Cron sample: `extra/upload.cron`

```
usage: upload [-h] [-p PATH] -H HOST [-U USER]

upload API submission files

optional arguments:
  -h, --help            show this help message and exit
  -p PATH, --path PATH  path to data storage
  -H HOST, --host HOST  upstream hostname
  -U USER, --user USER  upstream user credential
```

This script will automatically upload API submission files, c.f. `darc.submit`, using *`curl(1)`*. The `--user` option is supplied for the same option of *`curl(1)`*.

Important: As the `darc.submit.save_submit()` is categorising saved API submission files by its actual date, the script is also uploading such files by the saved dates. Therefore, as the *`cron(8)`* sample suggests, the script should better be run everyday *slightly after* **12:00 AM** (*0:00* in 24-hour format).

9.3 Remove Repeated Lines

File location `extra/uniq.py`

This script works the same as *`uniq(1)`*, except it filters one input line at a time without putting pressure onto memory utilisation.

9.4 Redis Clinic

File location

- Entry point: `extra/clinic.py`
- Helper script: `extra/clinic.lua`
- Cron sample: `extra/clinic.cron`

```
usage: clinic [-h] -r REDIS [-f FILE] [-t TIMEOUT] ...

memory clinic for Redis

positional arguments:
  services              name of services

optional arguments:
  -h, --help            show this help message and exit
  -r REDIS, --redis REDIS
                        URI to the Redis server
  -f FILE, --file FILE  path to compose file
  -t TIMEOUT, --timeout TIMEOUT
                        shutdown timeout in seconds
```

Since Redis may take more and more memory as the growth of crawled data and task queues, this script will truncate the Redis task queues (`queue_requests & ``queue_selenium`), as well as the corresponding `pickle` caches of `darc.link.Link`.

Note: We used Lua scrpit to slightly accelerate the whole procedure, as it may bring burden to the host server if running through Redis client.

CHAPTER

TEN

RATIONALE

There are two types of *workers*:

- `crawler` – runs the `darc.crawl.crawler()` to provide a fresh view of a link and test its connectability
- `loader` – run the `darc.crawl.loader()` to provide an in-depth view of a link and provide more visual information

The general process can be described as following for *workers* of `crawler` type:

1. `process_crawler()`: obtain URLs from the `requests` link database (c.f. `load_requests()`), and feed such URLs to `crawler()`.

   Note: If *FLAG_MP* is `True`, the function will be called with *multiprocessing* support; if *FLAG_TH* if `True`, the function will be called with *multithreading* support; if none, the function will be called in single-threading.

2. `crawler()`: parse the URL using `parse_link()`, and check if need to crawl the URL (c.f. *PROXY_WHITE_LIST*, *PROXY_BLACK_LIST*, *LINK_WHITE_LIST* and *LINK_BLACK_LIST*); if true, then crawl the URL with `requests`.

   If the URL is from a brand new host, *darc* will first try to fetch and save `robots.txt` and sitemaps of the host (c.f. `save_robots()` and `save_sitemap()`), and extract then save the links from sitemaps (c.f. `read_sitemap()`) into link database for future crawling (c.f. `save_requests()`). Also, if the submission API is provided, `submit_new_host()` will be called and submit the documents just fetched.

   If `robots.txt` presented, and *FORCE* is `False`, *darc* will check if allowed to crawl the URL.

   Note: The root path (e.g. `/` in https://www.example.com/) will always be crawled ignoring `robots.txt`.

   At this point, *darc* will call the customised hook function from *darc.sites* to crawl and get the final response object. *darc* will save the session cookies and header information, using `save_headers()`.

   Note: If `requests.exceptions.InvalidSchema` is raised, the link will be saved by `save_invalid()`. Further processing is dropped.

   If the content type of response document is not ignored (c.f. *MIME_WHITE_LIST* and *MIME_BLACK_LIST*), `submit_requests()` will be called and submit the document just fetched.

   If the response document is HTML (`text/html` and `application/xhtml+xml`), `extract_links()` will be called then to extract all possible links from the HTML document and save such links into the database (c.f. `save_requests()`).

And if the response status code is between `400` and `600`, the URL will be saved back to the link database (c.f. `save_requests()`). If **NOT**, the URL will be saved into `selenium` link database to proceed next steps (c.f. `save_selenium()`).

The general process can be described as following for *workers* of `loader` type:

1. `process_loader()`: in the meanwhile, *darc* will obtain URLs from the `selenium` link database (c.f. `load_selenium()`), and feed such URLs to `loader()`.

   Note: If *FLAG_MP* is `True`, the function will be called with *multiprocessing* support; if *FLAG_TH* if `True`, the function will be called with *multithreading* support; if none, the function will be called in single-threading.

2. `loader()`: parse the URL using `parse_link()` and start loading the URL using `selenium` with Google Chrome.

   At this point, *darc* will call the customised hook function from *darc.sites* to load and return the original `WebDriver` object.

   If successful, the rendered source HTML document will be saved, and a full-page screenshot will be taken and saved.

   If the submission API is provided, `submit_selenium()` will be called and submit the document just loaded.

   Later, `extract_links()` will be called then to extract all possible links from the HTML document and save such links into the `requests` database (c.f. `save_requests()`).

Important: For more information about the hook functions, please refer to the *customisation* documentations.

CHAPTER

ELEVEN

INSTALLATION

Note: *darc* supports Python all versions above and includes **3.6**. Currently, it only supports and is tested on Linux (*Ubuntu 18.04*) and macOS (*Catalina*).

When installing in Python versions below **3.8**, *darc* will use `walrus` to compile itself for backport compatibility.

```
pip install darc
```

Please make sure you have Google Chrome and corresponding version of Chrome Driver installed on your system.

Important: Starting from version **0.3.0**, we introduced Redis for the task queue database backend.

Since version **0.6.0**, we introduced relationship database storage (e.g. MySQL, SQLite, PostgreSQL, etc.) for the task queue database backend, besides the Redis database, since it can be too much memory-costly when the task queue becomes vary large.

Please make sure you have one of the backend database installed, configured, and running when using the *darc* project.

However, the *darc* project is shipped with Docker and Compose support. Please see *Docker Integration* for more information.

Or, you may refer to and/or install from the Docker Hub repository:

```
docker pull jsnbzh/darc[:TAGNAME]
```

or GitHub Container Registry, with more updated and comprehensive images:

```
docker pull ghcr.io/jarryshaw/darc[:TAGNAME]
# or the debug image
docker pull ghcr.io/jarryshaw/darc-debug[:TAGNAME]
```

CHAPTER

TWELVE

USAGE

Important: Though simple CLI, the *darc* project is more configurable by environment variables. For more information, please refer to the *environment variable configuration* documentations.

The *darc* project provides a simple CLI:

```
usage: darc [-h] [-v] -t {crawler,loader} [-f FILE] ...

the darkweb crawling swiss army knife

positional arguments:
  link                  links to craw

optional arguments:
  -h, --help            show this help message and exit
  -v, --version         show program's version number and exit
  -t {crawler,loader}, --type {crawler,loader}
                        type of worker process
  -f FILE, --file FILE  read links from file
```

It can also be called through module entrypoint:

```
python -m python-darc ...
```

Note: The link files can contain **comment** lines, which should start with `#`. Empty lines and comment lines will be ignored when loading.

CHAPTER

THIRTEEN

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

d

darc, 13
darc.error, 28
darc.model, 31
darc.model.tasks, 31
darc.model.web, 31
darc.proxy, 15
darc.sites, 22

INDEX

Symbols

_BaseException, 30
_BaseWarning, 30
__init__() (*darc.error.LinkNoReturn method*), 29

A

API_NEW_HOST, 15
API_REQUESTS, 15
API_RETRY, 15
API_SELENIUM, 15
APIRequestFailed, 29

C

CHROME_BINARY_LOCATION, 15

D

darc
    module, 13
darc.const.CHECK (*built-in variable*), 24
darc.const.CHECK_NG (*built-in variable*), 24
darc.const.CWD (*built-in variable*), 24
darc.const.DARC_CPU (*built-in variable*), 24
darc.const.DARC_USER (*built-in variable*), 24
darc.const.DARC_WAIT (*built-in variable*), 26
darc.const.DB (*built-in variable*), 25
darc.const.DEBUG (*built-in variable*), 23
darc.const.FLAG_DB (*built-in variable*), 25
darc.const.FLAG_MP (*built-in variable*), 24
darc.const.FLAG_TH (*built-in variable*), 24
darc.const.FORCE (*built-in variable*), 24
darc.const.LINK_BLACK_LIST (*built-in variable*), 27
darc.const.LINK_FALLBACK (*built-in variable*), 27
darc.const.LINK_WHITE_LIST (*built-in variable*), 27
darc.const.MIME_BLACK_LIST (*built-in variable*), 27
darc.const.MIME_FALLBACK (*built-in variable*), 27
darc.const.MIME_WHITE_LIST (*built-in variable*), 27
darc.const.PATH_DB (*built-in variable*), 25
darc.const.PATH_ID (*built-in variable*), 25
darc.const.PATH_LN (*built-in variable*), 25
darc.const.PATH_MISC (*built-in variable*), 25
darc.const.PROXY_BLACK_LIST (*built-in variable*), 28
darc.const.PROXY_FALLBACK (*built-in variable*), 28
darc.const.PROXY_WHITE_LIST (*built-in variable*), 27
darc.const.REBOOT (*built-in variable*), 23
darc.const.REDIS (*built-in variable*), 25
darc.const.ROOT (*built-in variable*), 24
darc.const.SE_EMPTY (*built-in variable*), 26
darc.const.SE_WAIT (*built-in variable*), 26
darc.const.TIME_CACHE (*built-in variable*), 26
darc.const.VERBOSE (*built-in variable*), 24
darc.db.BULK_SIZE (*built-in variable*), 13
darc.db.LOCK_TIMEOUT (*built-in variable*), 14
darc.db.MAX_POOL (*built-in variable*), 14
darc.db.REDIS_LOCK (*built-in variable*), 14
darc.db.RETRY_INTERVAL (*built-in variable*), 14
darc.error
    module, 28
darc.model
    module, 31
darc.model.tasks
    module, 31
darc.model.web
    module, 31
darc.parse.URL_PAT (*built-in variable*), 13
darc.proxy
    module, 15
darc.proxy.bitcoin.LOCK (*built-in variable*), 15
darc.proxy.bitcoin.PATH (*built-in variable*), 15
darc.proxy.data.PATH (*built-in variable*), 16
darc.proxy.ed2k.LOCK (*built-in variable*), 16
darc.proxy.ed2k.PATH (*built-in variable*), 16
darc.proxy.freenet._FREENET_ARGS (*built-in variable*), 17
darc.proxy.freenet._FREENET_BS_FLAG (*built-in variable*), 17

darc.proxy.freenet._FREENET_PROC (*built-in variable*), 17
darc.proxy.freenet._MNG_FREENET (*built-in variable*), 17
darc.proxy.freenet.BS_WAIT (*built-in variable*), 16
darc.proxy.freenet.FREENET_ARGS (*built-in variable*), 16
darc.proxy.freenet.FREENET_PATH (*built-in variable*), 16
darc.proxy.freenet.FREENET_PORT (*built-in variable*), 16
darc.proxy.freenet.FREENET_RETRY (*built-in variable*), 16
darc.proxy.i2p._I2P_ARGS (*built-in variable*), 18
darc.proxy.i2p._I2P_BS_FLAG (*built-in variable*), 18
darc.proxy.i2p._I2P_PROC (*built-in variable*), 18
darc.proxy.i2p._MNG_I2P (*built-in variable*), 18
darc.proxy.i2p.BS_WAIT (*built-in variable*), 17
darc.proxy.i2p.I2P_ARGS (*built-in variable*), 17
darc.proxy.i2p.I2P_PORT (*built-in variable*), 17
darc.proxy.i2p.I2P_REQUESTS_PROXY (*built-in variable*), 17
darc.proxy.i2p.I2P_RETRY (*built-in variable*), 17
darc.proxy.i2p.I2P_SELENIUM_PROXY (*built-in variable*), 17
darc.proxy.irc.LOCK (*built-in variable*), 18
darc.proxy.irc.PATH (*built-in variable*), 18
darc.proxy.LINK_MAP (*in module darc.proxy*), 22
darc.proxy.magnet.LOCK (*built-in variable*), 18
darc.proxy.magnet.PATH (*built-in variable*), 18
darc.proxy.mail.LOCK (*built-in variable*), 19
darc.proxy.mail.PATH (*built-in variable*), 18
darc.proxy.null.LOCK (*built-in variable*), 19
darc.proxy.null.PATH (*built-in variable*), 19
darc.proxy.script.LOCK (*built-in variable*), 19
darc.proxy.script.PATH (*built-in variable*), 19
darc.proxy.tel.LOCK (*built-in variable*), 19
darc.proxy.tel.PATH (*built-in variable*), 19
darc.proxy.tor._MNG_TOR (*built-in variable*), 21
darc.proxy.tor._TOR_BS_FLAG (*built-in variable*), 21
darc.proxy.tor._TOR_CONFIG (*built-in variable*), 21
darc.proxy.tor._TOR_CTRL (*built-in variable*), 21
darc.proxy.tor._TOR_PROC (*built-in variable*), 21
darc.proxy.tor.BS_WAIT (*built-in variable*), 20
darc.proxy.tor.TOR_CFG (*built-in variable*), 20
darc.proxy.tor.TOR_CTRL (*built-in variable*), 20
darc.proxy.tor.TOR_PASS (*built-in variable*), 20
darc.proxy.tor.TOR_PORT (*built-in variable*), 20
darc.proxy.tor.TOR_REQUESTS_PROXY (*built-in variable*), 19
darc.proxy.tor.TOR_RETRY (*built-in variable*), 20
darc.proxy.tor.TOR_SELENIUM_PROXY (*built-in variable*), 20
darc.proxy.zeronet._MNG_ZERONET (*built-in variable*), 22
darc.proxy.zeronet._ZERONET_ARGS (*built-in variable*), 22
darc.proxy.zeronet._ZERONET_BS_FLAG (*built-in variable*), 22
darc.proxy.zeronet._ZERONET_PROC (*built-in variable*), 22
darc.proxy.zeronet.BS_WAIT (*built-in variable*), 21
darc.proxy.zeronet.ZERONET_ARGS (*built-in variable*), 21
darc.proxy.zeronet.ZERONET_PATH (*built-in variable*), 21
darc.proxy.zeronet.ZERONET_PORT (*built-in variable*), 21
darc.proxy.zeronet.ZERONET_RETRY (*built-in variable*), 21
darc.save._SAVE_LOCK (*built-in variable*), 13
darc.selenium.BINARY_LOCATION (*built-in variable*), 15
darc.sites
    module, 22
darc.sites.SITEMAP (*in module darc.sites*), 23
darc.submit.API_NEW_HOST (*built-in variable*), 15
darc.submit.API_REQUESTS (*built-in variable*), 15
darc.submit.API_RETRY (*built-in variable*), 15
darc.submit.API_SELENIUM (*built-in variable*), 15
darc.submit.PATH_API (*built-in variable*), 14
darc.submit.SAVE_DB (*built-in variable*), 14
DARC_BULK_SIZE, 13
DARC_CHECK, 24
DARC_CHECK_CONTENT_TYPE, 24
DARC_CPU, 24
DARC_DEBUG, 24
DARC_FORCE, 24
DARC_LOCK_TIMEOUT, 14
DARC_MAX_POOL, 14
DARC_MULTIPROCESSING, 24
DARC_MULTITHREADING, 24
DARC_REBOOT, 23, 60
DARC_REDIS_LOCK, 14

DARC_RETRY, 14
DARC_SAVE, 37
DARC_SAVE_REQUESTS, 37
DARC_SAVE_SELENIUM, 37
DARC_URL_PAT, 13
DARC_USER, 25
DARC_VERBOSE, 24
DARC_WAIT, 26
DatabaseOperaionFailed, 29

E

environment variable
    API_NEW_HOST, 15, 40
    API_REQUESTS, 15, 40
    API_RETRY, 15, 39
    API_SELENIUM, 15, 40
    CHROME_BINARY_LOCATION, 15, 37
    DARC_BULK_SIZE, 13, 35
    DARC_CHECK, 24, 33
    DARC_CHECK_CONTENT_TYPE, 24, 34
    DARC_CPU, 24, 34
    DARC_DEBUG, 24, 33
    DARC_FORCE, 24, 33
    DARC_FREENET, 43
    DARC_I2P, 41
    DARC_LOCK_TIMEOUT, 14
    DARC_MAX_POOL, 14, 35
    DARC_MULTIPROCESSING, 24, 34
    DARC_MULTITHREADING, 24, 34
    DARC_REBOOT, 23, 33, 60
    DARC_REDIS_LOCK, 14
    DARC_RETRY, 14
    DARC_SAVE, 36, 37
    DARC_SAVE_REQUESTS, 37
    DARC_SAVE_SELENIUM, 37
    DARC_TOR, 40
    DARC_URL_PAT, 13, 34
    DARC_USER, 25, 34
    DARC_VERBOSE, 24, 33
    DARC_WAIT, 26, 36
    DARC_ZERONET, 42
    DB_URL, 35
    FREENET_ARGS, 43
    FREENET_PATH, 43
    FREENET_PORT, 43
    FREENET_RETRY, 43
    FREENET_WAIT, 43
    I2P_ARGS, 42
    I2P_PORT, 41
    I2P_RETRY, 41
    I2P_WAIT, 41
    LINK_BLACK_LIST, 27, 38
    LINK_FALLBACK, 27, 38
    LINK_WHITE_LIST, 27, 38
    LOCK_TIMEOUT, 35
    MIME_BLACK_LIST, 27, 38
    MIME_FALLBACK, 27, 39
    MIME_WHITE_LIST, 27, 38
    PATH_DATA, 25, 35
    PROXY_BLACK_LIST, 28, 39
    PROXY_FALLBACK, 28, 39
    PROXY_WHITE_LIST, 27, 39
    REDIS_LOCK, 36
    REDIS_URL, 25, 35
    RETRY_INTERVAL, 36
    SAVE_DB, 15, 39
    SE_WAIT, 26, 37
    TIME_CACHE, 26, 37
    TOR_CFG, 41
    TOR_CTRL, 40
    TOR_PASS, 40
    TOR_PORT, 40
    TOR_RETRY, 41
    TOR_WAIT, 41
    ZERONET_ARGS, 43
    ZERONET_PATH, 42
    ZERONET_PORT, 42
    ZERONET_RETRY, 42
    ZERONET_WAIT, 42

F

FreenetBootstrapFailed, 29

G

get_lock() (*in module darc.const*), 23

H

HookExecutionFailed, 29

I

I2PBootstrapFailed, 29

L

LINK_BLACK_LIST, 27
LINK_FALLBACK, 27
LINK_WHITE_LIST, 27
LinkNoReturn, 29
LockWarning, 29

M

MIME_BLACK_LIST, 27
MIME_FALLBACK, 27
MIME_WHITE_LIST, 27
module
    darc, 13
    darc.error, 28
    darc.model, 31
    darc.model.tasks, 31

darc.model.web, 31
darc.proxy, 15
darc.sites, 22

P

PATH_DATA, 25
PROXY_BLACK_LIST, 28
PROXY_FALLBACK, 28
PROXY_WHITE_LIST, 27

R

REDIS_URL, 25
RedisCommandFailed, 29
render_error() (*in module darc.error*), 30

S

SAVE_DB, 15
SE_WAIT, 26
SiteNotFoundWarning, 29

T

TIME_CACHE, 26
TorBootstrapFailed, 29
TorRenewFailed, 30

U

UnsupportedLink, 30
UnsupportedPlatform, 30
UnsupportedProxy, 30

W

WorkerBreak, 30

Z

ZeroNetBootstrapFailed, 30