twikiget Documentation

Release 0.0.1.dev20190306

CERN Analysis Preservation

Jul 18, 2019

Contents

twikiget is a tool to download twiki pages and archive them in .warc format. It uses wget underneath and so it includes all its downloading features.

CHAPTER 1

Introduction

1.1 About

twikiget is a tool to download twiki pages and archive them in .warc format. It uses wget underneath and so it includes all its downloading features.

1.2 Features

- download and archive specific TWiki page and all its attachments
- create WARC files for long-term preservation purposes
- save local cache for faster and periodic reprocessing
- (planned) extract specific metadata from TWiki document markup according to configurable templates

1.3 Useful links

- documentation
- releases
- known issues
- source code

CHAPTER 2

User Guide

2.1 Install twikiget

If you are interested in archiving twikis using twikiget, all you need to install is the `twikiget`, ideally in a new virtual environment:

```
$ # create new virtual environment
$ virtualenv ~/.virtualenvs/twikiget
$ source ~/.virtualenvs/twikiget/bin/activate
$ # install twikiget
$ pip install twikiget
```

2.2 Basic usage

```
$ # download twiki
$ twikiget archive https://twiki.cern.ch/twiki/bin/view/Main/ZhuTopAnalysis
$ ls
ZhuTopAnalysis.warc cache
$ # once the twiki is archived we can list it's contents:
$ twikiget list ZhuTopAnalysis.warc
$ ...
$ # we can also view the raw content of each file:
$ twikiget view ZhuTopAnalysis.warc https://twiki.cern.ch/twiki/bin/view/Main/
↪ZhuTopAnalysis
$ ...
```

2.3 CLI API

2.3.1 archive

Archive a TWiki page with attachments into a WARC archive.

Raw archived files are also saved to a directory specified in *directory-prefix* option (default=./cache).

Options passed in *wget-options* will overwrite the twikiget defaults, and should be used with caution.

```
archive [OPTIONS] URL
```

Options

--wget-options <wget_options>
: additional options to pass to wget

-o, --out-warc-file <out_warc_file>
: output file name for a WARC file

-P, --directory-prefix <directory_prefix>
: output folder for raw files

Arguments

URL
: Required argument

2.3.2 list

List files in the WARC archive.

The list can be filtered by the HTTP Content Type, and exported as json if needed.

Note that *content-type* option can be a full name of a type or, to search broader, just the first part of it e.g. *text/css* and *text*

Examples:

```
$ twikiget list ExampleTwiki.warc

$ twikiget list ExampleTwiki.warc --content-type=text/html

$ twikiget list ExampleTwiki.warc --content-type=text

$ twikiget list ExampleTwiki.warc --json
```

```
list [OPTIONS] WARC_FILE
```

Options

--json
: Get output in JSON format.

--content-type <filter_content_type>
Filter files in an archive by content_type. It can be either full version or just a begging of type name

Arguments

WARC_FILE
Required argument

2.3.3 view

View raw content of one of the files in the WARC archive.

View command is usefull to inspect contents of one file from the archive. It can be used with a pipe or a stream to view the file in a web-browser or other suitable program. FILE-URI argument can be copied form the outputs of *twikiget list*.

Examples:

```
$ twikiget view ExampleTwiki.warc https://example.com/twiki?raw=on
$ twikiget view ExampleTwiki.warc http://example.com/style.css

$ twikiget view ExampleTwiki.warc http://example.com/img.png > img.png
```

```
view [OPTIONS] WARC_FILE FILE_URI
```

Arguments

WARC_FILE
Required argument

FILE_URI
Required argument

CHAPTER 3

Contributing

3.1 Issues

Bug reports, feature requests, and other contributions are welcome. If you find a demonstrable problem that is caused by the twikiget code, please:

1. Search for already reported problems.
2. Check if the issue has been fixed or is still reproducible on the latest *master* branch.
3. Create an issue, ideally with **a test case**.

3.2 Pull requests

If you create a feature branch, you can run the tests to ensure that everything is operating correctly:

```
$ ./run-tests.sh
```

Each pull request should preserve or increase code coverage.

3.3 Kanban

We are using Kanban technique for keeping track of ongoing tasks. Please see our Kanban board and look for issues that are labelled as "Ready".

CHAPTER 4

Changes

4.1 Version 0.1.0 (2019-MM-DD)

- Initial public release.

Please beware

Please note that twikiget is in an early alpha stage of its development. The developer preview releases are meant for early adopters and testers. Please don't rely on released versions for any production purposes yet.

CHAPTER 5

License

MIT License

Copyright (C) 2019 CERN.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

In applying this license, CERN does not waive the privileges and immunities granted to it by virtue of its status as an Intergovernmental Organization or submit itself to any jurisdiction.

CHAPTER 6

Authors

The list of contributors in alphabetical order:

- Jan Okraska
- Tibor Simko

Index

Symbols

A

F

L

U

V

W