Crawler Documentation

Release 1.0

Andrey

May 28, 2016

Contents

1 Installation 3

2 Support 5

3 Cookbook 7
3.1 Crawl a web page 7
3.2 Crawl a page slowly 7
3.3 Crawl only your blog 7

4 Crawler Python API 9

5 Command Line Options 11

6 Indices and tables 13

Contents:

User Guide

CHAPTER 1

Installation

At the command line:

```
easy_install crawler
```

Or, if you have pip installed:

```
pip install crawler
```

CHAPTER 2

Support

The easiest way to get help with the project is to join the #crawler channel on Freenode. We hang out there and you can get real-time help with your projects. The other good way is to open an issue on Github.

The mailing list is also available for support.

- Freenode
- Github

CHAPTER 3

Cookbook

3.1 Crawl a web page

The most simple way to use our program is with no arguments. Simply run:

```
python main.py -u <url>
```

to crawl a webpage.

3.2 Crawl a page slowly

To add a delay to your crawler, use *-d*:

```
python main.py -d 10 -u <url>
```

This will wait 10 seconds between page fetches.

3.3 Crawl only your blog

You will want to use the *-i* flag, which while ignore URLs matching the passed regex:

```
python main.py -i "^blog" -u <url>
```

This will only crawl pages that contain your blog URL.

Programmer Reference

CHAPTER 4

Crawler Python API

Getting started with Crawler is easy. The main class you need to care about is crawler.main.Crawler

```
>>> should_ignore(['blog/$'], 'http://ericholscher.com/blog/')
True
```

```
>>> should_ignore(['home'], 'http://ericholscher.com/blog/')
True
```

```
>>> log('http://ericholscher.com/blog/', 200)
OK: 200 http://ericholscher.com/blog/
```

```
>>> log('http://ericholscher.com/blog/', 500)
ERR: 500 http://ericholscher.com/blog/
```

```
>>> log('http://ericholscher.com/blog/', 500)
OK: 500 http://ericholscher.com/blog/
```

CHAPTER 5

Command Line Options

These flags allow you to change the behavior of Crawler. Check out how to use them in the Cookbook.

-d <sec>, **--delay** <sec>

Use a delay in between page fetchs so we don't overwhelm the remote server. Value in seconds.

Default: 1 second

-i <regex>, **--ignore** <regex>

Ignore pages that match a specific pattern.

Default: None

CHAPTER 6

Indices and tables

- genindex
- modindex
- search

Index

Symbols

-d <sec>, –delay <sec>
    command line option, 11

-i <regex>, –ignore <regex>
    command line option, 11

C

command line option
    -d <sec>, –delay <sec>, 11
    -i <regex>, –ignore <regex>, 11