bg.crawler Documentation

Release 0.2

Andreas Jung

March 27, 2015

Contents

`bg.crawler` is a command-line frontend for feeding a tree of files (a directory) into a Solr for indexing.

CHAPTER 1

Requirements

- Python 2.6 or Python 2.7 (no support for Python 3)
- `curl`

CHAPTER 2

Installation

- use `easy_install bg.crawler` - this should install a script `solr-crawler` inside the `bin` folder of your Python installation. You are strongly encouraged to use `virtualenv` for creating a virtualized Python environment.

CHAPTER 3

Usage

Command line options:

```
usage: solr-crawler [-h] [--solr-url SOLR_URL]
                    [--render-base-url RENDER_BASE_URL]
                    [--max-depth MAX_DEPTH] [--commit-after COMMIT_AFTER]
                    [--tag TAG] [--clear-all] [--optimize] [--guess-encoding]
                    [--clear-tag SOLR_CLEAR_TAG] [--verbose] [--no-type-check]
                    <directory>

A command-line crawler for importing all files within a directory into Solr

positional arguments:
  <directory>           Directory to be crawled

optional arguments:
  -h, --help            show this help message and exit
  --solr-url SOLR_URL, -u SOLR_URL
                        SOLR server URL
  --render-base-url RENDER_BASE_URL, -r RENDER_BASE_URL
                        Base URL for server delivering crawled content
  --max-depth MAX_DEPTH, -d MAX_DEPTH
                        maximum folder depth
  --commit-after COMMIT_AFTER, -C COMMIT_AFTER
                        Solr commit after N documents
  --tag TAG, -t TAG     Solr import tag
  --clear-all, -c       Clear the Solr indexes before crawling
  --optimize, -O        Optimize Solr index after import
  --guess-encoding, -g  Guess encoding of input data
  --clear-tag SOLR_CLEAR_TAG
                        Remove all items from Solr indexed tagged with the
                        given tag
  --verbose, -v         Verbose logging
  --no-type-check, -n   Do not apply internal extension filter while crawling

Have fun!
```

- `--solr-url` defines the URL of the SOLR server
- **`--render-base-url` can be specified in order specify an URL prefix in order** to calculate the value of the `renderurl` field within Solr. The value of `renderurl` is the concatenation of the value for `render-base-url` and the relative path of the crawled file to the crawler start directory. This is option is useful for generating a link using the `renderurl` if the file is served through a given web server (by its URL).

- `--max-depth` limits the crawler to a given folder depth
- **`--commit-after` can be used to specify a numeric value to** import the documents into batches with a Solr commit operation after each batch instead of committing after each individual document.
- `--tag` will tag the imported document(s) with a string (this may be useful importing different document sources into Solr while supporting the option to filter by tag at query time)
- `--clear-all` clear the complete Solr index before running the import
- `--clear-tag` remove all documents with the given tag before running the import
- `--verbose` enable extensive logging
- `--no-type-check` if set: do not apply any type check filtering but instead pass all file types to Solr

CHAPTER 4

Internals

- uses the `python-magic` module to determine the mimetype of files to be imported
- currently deals with HTML and plain text files
- HTML files are currently parsed internally and converted to plain text

CHAPTER 5

Sourcecode

https://github.com/zopyx/bg.crawler

CHAPTER 6

Bug tracker

https://github.com/zopyx/bg.crawler/issues

CHAPTER 7

Solr setup

You can use the buildout configuration from

https://raw.github.com/zopyx/bg.crawler/master/solr-3.5.cfg

as an example how to setup a Solr instance for using `bg.crawler`.

It is important that the following field type definition is available within your Solr instance:

```
index =
    name:text             type:text    stored:true
    name:title            type:text    stored:true
    name:created          type:date    stored:true required:true
    name:modified         type:date    stored:true
    name:filesize         type:integer stored:true
    name:mimetype         type:string  stored:true
    name:id               type:string  stored:true required:true
    name:relpath          type:string  stored:true
    name:fullpath         type:string  stored:true
    name:renderurl        type:string  stored:true
    name:tag              type:string  stored:true
```

After running `buildout` you can start the Solr instance using:

```
bin/solr-instance fg|start
```

CHAPTER 8

Licence

`bg.crawler` is published under the GNU Public Licence V2 (GPL 2)

CHAPTER 9

Author

ZOPYX Ltd.
Charlottenstr. 37/1
D-72070 Tuebingen
Germany
info@zopyx.com
www.zopyx.com