ciur Documentation

Release 1.0.0

Andrei Danciuc

July 09, 2017

Contents

Ciur is a scrapper layer in development

Ciur is a lib because it has less black magic than a framework

It exports all scrapper related code into separate layer.

If you are annoyed by Spaghetti code, sql inside php and inline css inside html THEN you also are annoyed by xpath/css code inside crawler.

Ciur gives the taste of Lasagna code generally by enforcing encapsulation for scrapping layer.

It tries to not repeat the bad code.

CHAPTER 1

What does *Ciur* mean?

Ciur is Romanian for Sieve.

It fulfils the same purpose in the sense of being a “device for separating wanted elements from unwanted material”.

Python ciur API

```
>>> import ciur
>>> from ciur.shortcuts import pretty_parse_from_resources
>>> with ciur.open_file("example.org.ciur", __file__) as f:
...     print pretty_parse_from_resources(
...             f,
...             "http://example.org"
...     )
{
     "root": {
         "name": "Example Domain",
         "paragraph": "This domain is established to be used for illustrative examples in documents.
     }
 }
```

Samples of usage:

- Say Hello World in ciur language with http://www.example.org
- Container Docker + lambda amazon + Ciur combination for cuir
- Exchange money rates world wide parsers based on Ciur –> parsing world wide (40 sources, 4 country) currency exchange rates.
- https://bitbucket.org/ada/ciur.example.social –> parsing networking sites (such as Facebook, Linkedin, Xing ...) (not yet ready for open realease)

CHAPTER 2

For Developers:

- Local Python Virtual environment for cuir

CHAPTER 3

TODO:

- TODO: http://lybniz2.sourceforge.net/safeeval.html
- demo on cloud9
- build documentation on readthedocs
- http://lxml.de/lxmlhtml.html#parsing-html

  .cssselect(expr):

  .base_url:

Ciur Documentation

Install ciur

Lets assume that we using virtual env (see Python Virtual environment)

```
PIP=/opt/python-env/ciur_env/bin/pip
CIUR=/opt/python-env/ciur_env/bin/ciur
```

Install branch python3.6-ciur with pip

```
$ ${PIP} install "git+https://bitbucket.org/ada/python-ciur.git@python3.6-ciur#egg=ciur"
# or for contribution purposes
# ${PIP} install -e "/your/local/clone/of/ciur/branch"
...
Successfully installed cffi-1.4.2 ciur-0.1.2 cryptography-1.1.2
cssselect-0.9.1 enum34-1.1.2 html5lib-0.9999999 idna-2.0 ipaddress-1.0.16
lxml-3.5.0 ndg-httpsclient-0.4.0 pdfminer-20140328 pyOpenSSL-0.15.1
pyasn1-0.1.9 pycparser-2.14 pyparsing-2.0.7 python-dateutil-2.4.2
requests-2.9.1 six-1.10.0
...
```

Type *"Hello word"*

```
${CIUR} --url "http://example.org" --rules="https://bitbucket.org/ada/python-ciur/raw/python3.6-ciur/
```

Based on ciur rules:

```
$ curl "https://bitbucket.org/ada/python-ciur/raw/python3.6-ciur/docs/docker/example.org.ciur"
root `/html/body` +1
```

```
    name `.//h1/text()` +1
    paragraph `.//p/text()` +1
```

We are going to receive parsed data as json:

```
{
    "root": {
        "name": "Example Domain",
        "paragraph": "This domain is established to be used for illustrative
                      examples in documents. You may use this
                      domain in examples without prior coordination or
                     asking for permission."
    }
}
```

Python virtual environment

We will use only python version 3.5

Compile python it from source code

In case you don not have it, follow bellow instructions to compile it from source code.

```
#!/bin/bash
# script: compile_python

PYTHON_VERSION=3.6.1

cd /opt
wget --version > /dev/null || apt-get install -y wget # install wget in case is not present
wget -c "https://www.python.org/ftp/python/${PYTHON_VERSION}/Python-${PYTHON_VERSION}.tar.xz"

xz --version || apt-get install -y xz-utils  # install xz in case is not present
tar xf Python-${PYTHON_VERSION}.tar.xz
cd Python-${PYTHON_VERSION}/

gcc --version > /dev/null || apt-get install -y build-essential  # install xz in case is not present
apt-get install -y libssl-dev # ssl is required by PIP module

./configure
make
./python --version # should show Python ${PYTHON_VERSION}
```

Create python virtual environment

```
#!/bin/bash

sudo ${PYTHON_INTERPRETER_PATH}/python -m venv /opt/python3.6-ciur
```

Then use `/opt/python3.6-ciur/bin/python` as a default python interpreter in your IDE (f.e. PyCharm)

Install requirements

```
#!/bin/bash
# script: install_requirements

PYTHON_CIUR=/opt/python3.6-ciur/bin
${PYTHON_CIUR}/pip install --upgrade pip setuptools
apt-get install -y --force-yes $(curl "https://bitbucket.org/ada/python-ciur/raw/python3.6-ciur/requ

${PYTHON_CIUR}/pip install -r "https://bitbucket.org/ada/python-ciur/raw/python3.6-ciur/requirements-
```

Continuous Integration

travis-ci.com

Unfortunately travis do not support bitbucket see https://github.com/travis-ci/travis-ci/issues/667

magnum-ci.com

Dependencies installation:

sudo apt-get -y update sudo apt-get install -y python-pip libxml2-dev libxslt1-dev python-dev cython zlib1g-dev sudo pip install –upgrade setuptools sudo pip install –upgrade pip sudo pip install -r requirements-pip-dev.txt

Test suite commands:

python setup.py test

If you can't find the information you're looking for, have a look at the index or try to find it using the search function:

- genindex
- search