JobCrawler Documentation

Release 1772944

CROSS Solution

Sep 17, 2019

Contents

Autors	• Carsten Bleek
Contact	bleek@cross-solution.de
Revision	1772944
Date	Sep 17, 2019

CHAPTER 1

About

This document descibes, how to write Job Crawlers using scrapy. Sources of the documentation ist located at git@gitlab.cross-solution.de:scrapy/docs.git

By pushing to the master branch, rendering of the documentation is triggered.

CHAPTER 2

Installation

2.1 Installation of Python

We're using Python 3.x because Python 3 handles non English letters better due to native UTF-8 unicode support. Python 3 is the future of Python, and only old projects should maintain Python 2 compatability. We want to avoid installing or using Python 2 and modern versions of Ubuntu by default do not come with Python2. We only care about Python 3.x compatability for our crawlers.

```
sudo apt-get -y install build-essential python3-dev
wget -qO- https://bootstrap.pypa.io/get-pip.py | sudo -H python3
```

2.2 Installation of Scrapy

Once you have your Python3 dev environment setup the following will install Scrapy:

```
sudo -H pip3 install scrapy
```

2.3 Installation of Scrapyd

scrapyd provides an API to start, stop, schadule, etc. crawlers.

```
adduser --system --home /var/lib/scrapyd --gecos "scrapy" --no-create-home \
        --disabled-password --quiet scrapy
mkdir -p /var/lib/scrapyd \
         /var/log/scrapyd \
         /var/lib/scrapyd/eggs \
         /var/lib/scrapyd/dbs \
         /var/lib/scrapyd/items
chown scrapy:nogroup /var/log/scrapyd /var/lib/scrapyd \
         /var/lib/scrapyd/eggs \
         /var/lib/scrapyd/dbs \
         /var/lib/scrapyd/items
sudo apt-get -y install libjpeg-dev libfreetype6-dev zlib1g-dev libpng12-dev curl
pip install scrapyd
```

Next create a service file to start/stop scrapyd via systemd

```
root@scrapyd:/usr/lib/systemd/user# cat /lib/systemd/system/scrapyd.service
[Unit]
Description=Scrapyd service
After=network.target

[Service]
User=scrapy
Group=nogroup
WorkingDirectory=/var/lib/scrapyd
ExecStart=/usr/local/bin/scrapyd
Restart=always

[Install]
WantedBy=multi-user.target
```

Enable the service

```
root@scrapyd:~# systemctl enable scrapyd.service
Created symlink from /etc/systemd/system/multi-user.target.wants/scrapyd.service␣
↪to /lib/systemd/system/scrapyd.service.
```

The configuration of scrapyd is done in `/etc/scrapyd/scrapyd.conf`

```
root@scrapyd:~# cat /etc/scrapyd/scrapyd.conf
[scrapyd]
eggs_dir    = /var/lib/scrapyd/eggs
logs_dir    = /var/log/scrapyd/
items_dir   = /var/lib/scrapyd/item
jobs_to_keep = 5
dbs_dir     = dbs
max_proc    = 0
max_proc_per_cpu = 4
finished_to_keep = 100
poll_interval = 5.0
bind_address = 0.0.0.0
http_port   = 6800
debug       = off
runner      = scrapyd.runner
application = scrapyd.app.application
launcher    = scrapyd.launcher.Launcher
webroot     = scrapyd.website.Root

[services]
schedule.json     = scrapyd.webservice.Schedule
cancel.json       = scrapyd.webservice.Cancel
addversion.json   = scrapyd.webservice.AddVersion
listprojects.json = scrapyd.webservice.ListProjects
listversions.json = scrapyd.webservice.ListVersions
listspiders.json  = scrapyd.webservice.ListSpiders
delproject.json   = scrapyd.webservice.DeleteProject
delversion.json   = scrapyd.webservice.DeleteVersion
listjobs.json     = scrapyd.webservice.ListJobs
daemonstatus.json = scrapyd.webservice.DaemonStatus
```

CHAPTER 3

Guidelines

3.1 Creating a crawler

Crawlers are ordered by customers. So if a new crawler should be created, we start with an Issue in the create-new-scrapy-crawler Project. The Subject should contain the name of the crawler. This name should match the regular expression `/[a-z0-9-]+/`. The name should be used as name of the repository, name of the json file and name of the scrapy crawler.

Subject: create crawler <spidername> http://example.com

Developers are asked if they can do the task. Developers should assign themselve to the issue and start working. They should create a repository in the JobCrawler group. The new repository should be named <spidername>.

3.1.1 Example

```
root@scrapy-runner:~# scrapy startproject spidername
 New Scrapy project 'spidername', using template directory '/usr/local/lib/python3.
↪5/dist-packages/scrapy/templates/project', created in:
 /root/spidername

 You can start your first spider with:
     cd spidername
     scrapy genspider example example.com

 root@scrapy-runner:~/spidername# scrapy genspider spidername cross-solution.de
 Cannot create a spider with the same name as your project
```

Why spider cannot have the same name as the project? Please complete the example.

Please follow the PEP 8 Style Guide for Python Code

Please add a `.gitlab-ci.yml` to your code. Example

```
before_script:
   - export PATH=$PATH:/usr/local/bin
   - run-tests.sh ${CI_PROJECT_NAME}
```

This will test the code against the PEP 8 styleguide and execute the crawler on the test and deployment mashine by pushing changes to the master.

in addition create a `.gitignore` file:

```
*~
*.pyc

# ignore .idea and build directory
.idea
build
```

crawlers have to be version controlled by git. The location in our gitlab is: https://gitlab.cross-solution.de/scrapy/JobCrawler

3.2 Deployment

a Crawler is deployed by pushing changes to the master. The CI Feature of gitlab executes the crawler. All crawlwers are writung the structured data inti the local firectory /tmp/scrapy-json/

3.3 Format

all scrapy crawlers should create a json with the following format. The fieldnames are defined by the JobEntity of the YAWIK Project

`simple-import-schema.json`

```
{
    "$id": "https://scrapy-docs.yawik.org/build/html/_downloads/simple-import-schema.json",
    "$schema": "http://json-schema.org/draft-07/schema#",
    "description": "Defines a List oft Job Postings",
    "type": "object",
    "properties": {
        "jobs": {
            "description": "uniq identifier of the job posting",
            "type": "array",
            "items": {
                "type": "object",
                "properties": {
                    "id": {
                        "description": "uniq identifier of the job posting",
                        "type": "string"
                    },
                    "title": {
                        "description": "title of a job posting",
                        "type": "string"
                    },
                    "location": {
                        "description": "location of a job posting",
                        "type": "string"
                    },
                    "company": {
                        "description": "name of the company",
                        "type": "string"
                    },
                    "reference": {
                        "reference": "reference of the job posting used by the hiring organization",
                        "type": "string"
                    },
                    "contactEmail": {
```

```
                        "description": "email address for applications (if available)",
                        "type": "string"
                    },
                    "language": {
                        "description": "language of the job posting",
                        "type": "string"
                    },
                    "link": {
                        "description": "link to the detail page of the job posting",
                        "type": "string"
                    },
                    "datePublishStart": {
                        "type": "string",
                        "description": "date of the job posting (date format DD.MM.YYYY)"
                    },
                    "datePublishEnd": {
                        "description": "End date of the job posting (date format DD.MM.YYYY)",
                        "type": "string"
                    },
                    "logoRef": {
                        "description": "link to a logo of the hiring organization",
                        "type": "string"
                    },
                    "linkApply": {
                        "description": "link which references an application form",
                        "type": "string"
                    },
                    "classifications": {
                        "type": "object",
                        "properties": {
                            "professions": {
                                "type": "array",
                                "items": {
                                    "type": "string"
                                }
                            },
                            "industries": {
                                "type": "array",
                                "items": {
                                    "type": "string"
                                }
                            },
                            "employmentTypes": {
                                "type": "array",
                                "items": {
                                    "type": "string"
                                }
                            }
                        }
                    },
                    "templateValues": {
                        "type": "object",
                        "properties": {
                            "description": {
                                "description": "Introduction of the job advertisement. Usually a description of the company",
                                "type": "string"
                            },
                            "tasks": {
```

```
                                "description": "Description of the tasks.",
                                "type": "string"
                            },
                            "requirements": {
                                "description": "Description of the requirements or qualifications",
                                "type": "string"
                            },
                            "benefits": {
                                "description": "",
                                "type": "string"
                            },
                            "html": {
                                "type": "string"
                            }
                        }
                    }
                },
                "required": [
                    "id",
                    "title",
                    "link"
                ]
            }
        }
    }
}
```

Example:

```
{
   "jobs": [
       "id": Example-123/456,    // must not contain dots
       "title": "title of the job posting",
       "location": "location of the job posting",
       "company": "name of the hiring organization",
       "description":
       "reference": "reference of the job posting used by the hiring organization",
       "contactEmail": "email address for applications (if available)",
       "language": "language of the job posting",
       "link": "link to the detail page of the job posting",
       "datePublishStart": "date of the job posting (date format DD.MM.YYYY)",
       "datePublishEnd": "End date of the job posting (date format DD.MM.YYYY)",
       "logoRef": "link to a logo of the hiring organization",
       "linkApply": "link which references an application form",
       "classifications": {
           "professions": [
               "software-developer",
               "sales manager"
           ],
           "industries": [
               "banking",
               "IT"
           ],
           "employmentTypes": [
               "contract",
               "internship",
               "freelancer"
           ]
       },
       "templateValues":{
           "introduction": "<p></p>",
           "description": "<p>We're a good company<\/p>",
```

```
            "tasks":"<b>Your Tasks<\/b><ul><li>Task 1<\/li><li>Task2<\/li><\/ul>",
            "requirements":"<b>Qualifications<\/b><ul><li>requirement 1<\/li><li>
↪requirement 2<\/li<<\/ul>",
            "benefits":"<b>We offer<\/b><ul><li>offer 1<\/li><li>offer 2<\/li><\/ul>
↪",
            "boilerplate": "<p></p>"
            "html": "<p>complete html<\/p>"
        }
    ],[
        .....
    ]

}
```

field	
id	unique identifier. Must not contain dots '.'
title	title of the job posting
location	location of the job posting
link	link to the detail page of the job posting

The fields `id`, `title` and `link` are required. All other fields are optional. If the data can be crawled, put them into the described JSON format

The fields `datePublishStart` and `datePublishEnd` should be in the format DD.MM.YYYY

3.3.1 basis crawlers

A Basis crawler requires the fields required for a job list. These are:

- id
- title
- link
- location (if available in the overview)

A basic crawler is created to show that crawling is basically possible.

3.3.2 full crawler

A Full crawler should contain all data, which are a available using the job listing and the job detail page. Full crawlers are created when it is clear who pays for the work.

CHAPTER 4

Todo's

CHAPTER 5

Indices and tables

- genindex