Winit

Webscrapper for Windows

Installation guide

Libor Polčák, Tomáš Kocman

TARZAN project VI20172020062 document
Faculty of Information Technology, Brno University of Technology

Last change: April 29, 2020

Winit — Installation guide

Libor Polčák, Tomáš Kocman

Faculty of Information Technology, Brno University of Technology, e-mail: polcak@fit.vutbr.cz

This tool allows web page content scrapping and exporting the content as a compressed archive. The web crawl is performed using user-supplied regular expressions that may represent for example Torrent file names, Bitcoin wallets or keywords. Collected data may be used for law enforcement and other entitites, such as searching for information about a specific product or personal archive of web pages. The tool can be used in Microsoft Windows.

The aim of this document is to install the tool.

1 Project files

Download and extract the winit archive.

Launch command prompt.

Got to the `scrapitlite` directory of the extracted archive.

2 Python virtual environment

Install appropriate Windows installer of Python from https://www.python.org/downloads/windows/. The tool was tested with https://www.python.org/ftp/python/3.6.8/python-3.6.8-amd64.exe.

Do not forget to add Python to the PATH during install.

Create Python virtual environment (make sure that you are in the `scrapitlite` directory).

`python -m venv .`

Activate Python virtual environment.

```
cd Scripts
activate.bat
cd ..
```

Your command line prompt should be similar to:

`(scrapitlite) C:\Path\scrapitlite>`

3 C++ build tools

Download C++ build tools from https://www.microsoft.com/en-us/download/details.aspx?id=48159.

Install C++ build tools are installed in order to compile Python binary dependencies.

Download Microsoft Visual Studio from https://visualstudio.microsoft.com/downloads/ and install it. Make sure that you install Python development and native development tools, also install C++.

4 Installing dependencies

In the virtual environment in the command line, go to the folder with the downloaded winit archive (you will most probably need to run `cd ..`). Make sure that when running dir command, you see the file `requirements.txt`.

Install dependencies from `requirements.txt`:

```
python -m pip install -r requirements.txt
```

5 Run winit

Edit `scrapitlite\basicplatform\settings.py`, for example:

```
ALLOWED_DOMAINS = ["fit.vutbr.cz", "fit.vut.cz"]
START_URLS = ["https://www.fit.vut.cz"]
PATTERN = "Tarzan"
```

```
cd scrapitlite; scrapy crawl basic
```

See the results in `lemmitlite\output`.