

**Image Crawler and Data Cleaning Tool**

**User Guide and Setup Instruction**

**By Soon Yik Lee**

# **About this document**

**Intended audience**

This document is intended to be used by the tool user and others who want to test run the code for research or educational purpose. The source code is available for download at <https://github.com/LeeSoonYikUDE/SocialMediaCraweler>

**System requirements**

* Preferable wired network or stable wireless network
* Computer with the following non-tool related software installed and recommended hardware specification:
* Windows 10 64 bit
* 8 GB Ram and above
* GPU equipped
* Google Chrome

**NOTE:** For the image crawling tool, stable internet connection is essential to have a stable output, for data cleaning tool, the better performance the PC in terms of GPU, the more efficient the tool.

# **2. Overview of the tool**

**The Image Crawling Tool**

The Image Crawling Tool is a tool for crawling images from Facebook and Instagram. A Graphical User Interface will pop out when you run the tool, containing some fields that need to be fill before executing the crawling process.

When the tool executing the crawling process, it will utilize Chrome Web Driver to run the task in an automated way. In the end, the image from the webpage given will be crawled in specified directory, and long with an output CSV file.

The code is available at **ImageCrawler.py** file

<https://github.com/LeeSoonYikUDE/SocialMediaCraweler/blob/main/ImageCrawler.py>

**The Data Cleaning Tool**

The data cleaning tool is a tool specialized with the purpose of creating dataset containing only **German Meme** images. It contains 3 essential data cleaning functionalities:

1. **Duplication and near duplicate detection**

* Detect duplication or near duplication for the images. Option to remove it from directory once detected is available.
* **Non-German text detection**
* Detect whether the chosen image containing non-German text or having too many or too less words which exclude the image to be qualified as German meme. option to remove it from directory once detected is available.

1. **Plain Background detection**

* Detect the number of colors in the images, to determine whether it has only plain background without any object. Option to remove it from directory once detected is available.

1. **Image Normalization**

* To change the name and format of image in standard way

The code is available at DataCleaning.py file

<https://github.com/LeeSoonYikUDE/SocialMediaCraweler/blob/main/DataCleaning.py>

**3. Using the Image Crawling tool**

**Setting up Python**

If your PC is not already installed with Python, do that now. You can download Python package and run the executable file. The software package is available to download at <https://www.python.org/downloads/>

**Install Necessary Package**

When Python is ready to user, install the necessary libraries in order to run the code, refer to the command below:

pip install selenium

pip install PySimpleGUI

pip install wget

pip install pandas

pip install urllib3

pip install bs4

**Note:** If package missing error show up when running the code, look for the error details and install the necessary package. Optionally, you can run **pip install -r packagelist.txt** after download the packagelist.txt



**Setup Web Driver**

To setup the Web driver, first we need to know the version of Google Chrome browser being use in your system. You can check the chrome version by browsing <chrome://version/>

A picture containing scatter chart

Description automatically generated

As an example, the system have Chrome version of 96.0.4664.110 . With the version number, look for the webdriver for the version through <https://chromedriver.chromium.org/downloads> . Download the zip file, extract and paste the ChromeDriver.exe at C:\Users directory

Graphical user interface, website

Description automatically generated

**Note:** In most cases, web driver with the same version works and version after “96” can be neglected, look for the exact version when web driver related error come out when run the code.

**Login the Social Media website**

If the chosen account of Facebook and Instagram need to login the first time before running the code, to prevent verification when the account logged in to the brand-new machine. If account logged in using chrome browser without any issue, then it should now cause any issue during the login phase.

**Running the code**

If everything setup, when running the code, a GUI will pop up as shown below:

Graphical user interface, text, application, email

Description automatically generated

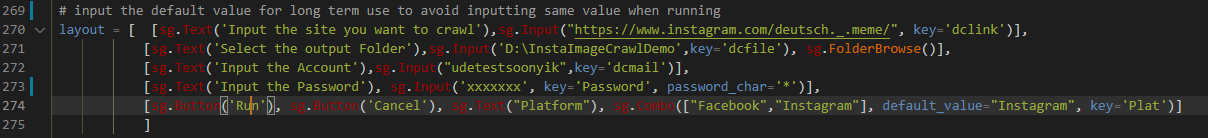
For testing purpose, you can try it with the websites shown below:

<https://www.facebook.com/Deutsch-Spa%C3%9F-104211548118278/>

<https://www.instagram.com/deutsch._.meme/>

For the account, it is suggested to not use your personal social media account, please register a new account solely for crawling purpose. There are 2 platforms to choose, which are Facebook and Instagram, input the Instagram page link if Instagram platform is chosen, and viceversa.

For long term use, it is recommended to input the default entry in the code, as shown below:



**Check The output**

If everything run smoothly, a window saying “Operation completed” will pop out, check the files in chosen directory, list of images and CSV should appear in the directory.

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, application, table, Excel

Description automatically generated

For video demo, please refer to <https://www.youtube.com/watch?v=SMa1PzgUPuU>

# **3. Using the Data Cleaning tool**

**Setting up Python**

If your PC is not already installed with Python, do that now. You can download Python package and run the executable file. The software package is available to download at <https://www.python.org/downloads/>

**Install Necessary Package**

When Python is ready to user, install the necessary libraries in order to run the code, refer to the command below:

pip install argparse

pip install imutils

pip install numpy

pip install scikit-image

pip install matplotlib

pip install opencv-python

pip install easyocr

pip install langdetect

pip install scipy

**Note:** If package missing error show up when running the code, look for the error details and install the necessary package.

**Running the code**

To run the code , in the command field of same directory, run the code with following command:

**python datacleaning.py -h**

it will show the option available to use the tool:

Text

Description automatically generated

Below is the description of each argument:

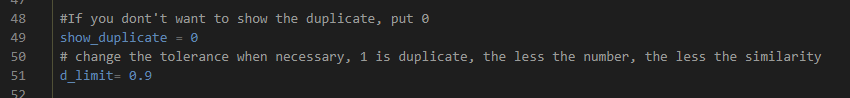
|  |  |  |  |
| --- | --- | --- | --- |
| Argument | Required | Description | Example |
| -h | optional | Show all available argument | datacleaning.py -h |
| -p PATH | Yes | The path of the images | datacleaning.py -p D:\test |
| -r | Yes | Remove image once detected, Y for yes and N for no | datacleaning.py -p D:\test -r N |
| -d | optional | Duplicate and near duplicate detection | datacleaning.py -p D:\test -r N -d |
| -l | optional | Non-German text detection | datacleaning.py -p D:\test -r N -l |
| -t | optional | Plain background check | datacleaning.py -p D:\test -r N -t |
| -f | optional | Normalize the image size and format | datacleaning.py -p D:\test -r N -t |

**Changing the value**

Some of the value and option can be modified in each functionalities according to needs. Below is the description of modification that can be made

1. **Duplication and near duplication**

On the **duplicate\_remove** function, there is 2 attributes to modify:



|  |  |
| --- | --- |
| Show\_duplicate | What will happen |
| 0 | It will not show the image comparison window |
| 1 | It will show the image comparison window |

The image comparison window looks like below

Graphical user interface, website

Description automatically generated

For **d\_limit** , the value set between **0 to 1**, the larger the value , the tool will accept the higher similarity duplication.

|  |  |
| --- | --- |
| d\_limit | What will happen |
| 0 <= x => 1 | The larger the value, higher tolerance the duplicate detection. When 0 is used, all images will count as duplication, when 1 is used, all images will not count as duplication. Recommended value is 0.9 |

1. **Language and word count detection**

On the main function section, there is an option to enable and disable GPU usage.

Text

Description automatically generated

For computer will only CPU, change the parameter to **gpu= false**, else change it to True.

|  |  |
| --- | --- |
| GPU | What will happen |
| True | GPU will be used when running language and word count detection |
| False | CPU will be used when running language and word count detection |

1. **Plain Background detection**

On the **only\_text\_remove** function, there is option to change value of number of colors found in images.

Text

Description automatically generated

The value of **cmax** determine the tolerance of image count as plain background. The higher the value, images with less color will count as plain background image

|  |  |
| --- | --- |
| cmax | What will happen |
| 0 <= x => 148 | The larger the value , the lesser the tolerance, choose 0 will let all images count as not plain background, set as 148 will count every image as plain background |

**Checking the output**

For demonstration purpose, you can use the DataCleaningDemo folder in the github.

From the terminal , you can check which image is detected as not qualified.

For the video instruction, refer <https://youtu.be/HynOQWbzD4c>