University of Groningen

Web Scraping Tool

Short Programming Project

User Guide

Version: 1.2

Author:
Keiko Angela Nicolasky(S3807452)

Supervisor:
Estefanía Talavera Martinez

rijksuniversiteit
groningen

Contents

1 Downloading the Tool

The tool is available to download from a GitHub repository. User can clone the repository from the link below:

```
https://github.com/keikoang/web-scraping-tool.git
```

This code is written in Python, and Python 3 is recommended to run it.

2 Installing Required Packages

Upon pulling the code from GitHub, user needs to install the required packages listed in `requirements.txt`. After running terminal on current working directory, user should enter the following command:

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

3 Entering Keywords/Usernames to be Downloaded

A `txt` file named `classes.txt` is provided when user pull the code from GitHub. In this `txt` file, user should put the keywords/usernames that he/she wants to download, separated by a newline. User can put a keyword that contain white-space character. For example 'cute ginger cat' in one line will be parsed into one keyword, 'cutegingercat'. User is not allowed to change the name of this file or move it to another directory.

Extra restriction is applied in case user wants to download a post from Instagram that contains two specific hashtags (both hashtags present in the post). The user then should only put those two hashtags in `classes.txt` file. For example, `classes.txt` should have this structure:

```
hashtag1
hashtag2
```

4 Choosing Web to Download Contents From Number of Contents to be Downloaded

After running `main.py`, user will be prompted this on terminal:

```
(1) Google
(2) Instagram
(3) Twitter
Enter (1), (2), or (3):
```

Simply enter number 1 or 2 in the terminal (without the parentheses) to indicate the desired web.
After that, the user will be asked to enter the number of samples he/she wants to download from the chosen web.

```
Enter number of samples to be downloaded:
```

Samples can be images, videos, or both. The entered number indicates how many samples per keyword/hashtag that will be downloaded. It is important to note that depending on the keywords, the actual numbers of downloaded samples may be less than the entered number. When downloading samples from Instagram, it may also be the case that the downloaded images or videos are more than the entered number. The reason behind this behaviour, is because Instagram scraper download posts. A post in Instagram can contain multiple images and videos.

For downloading samples that contain two specific hashtags (only available for Instagram), the entered number will be the number of samples that contain both hashtags. For example, a user wants to download 30 posts that has both cat and dog hashtags, then 30 should be entered instead of 15.

5 Google Image Scraper

Google Image scraper only downloads images from Google Image. After the user chose Google and entered the number of samples to be downloaded, the program will take some time to download the images.

The program will create a path `/database/google` and inside the google folder, there will be folder(s) named according to the keyword(s). User can find the images of each keywords in the the corresponding folder. The images are named in `keyword_index` pattern, where keyword is the keyword for the image, and index is a unique number indicating the image ranging from 1 up to `n`, with `n` being the number of samples to be downloaded.

Additionally, user can find a descriptor file in each of keyword folders. A descriptor file describes the following properties of the downloaded images: caption, width, height. The line number of each descriptor corresponds to index of downloaded image.

Below is an example of how the directories look like when a user downloads 3 images related to cat:

```
database
--google
----cat
------cat_1.jpg
------cat_2.jpg
------cat_2.jpg
------google_cat_descriptor.csv
----google_log
--instagram
--twitter
```

6 Instagram Scraper

Beside downloading images, Instagram scraper has an extra feature, which is downloading videos. After the user chose Instagram and entered number of samples to be downloaded, the user will be asked to indicate if he/she wants to download posts in certain time period.

```
Download posts in certain time period?
(1) Yes
(2) No
Enter (1) or (2):
```

If a user wants to download posts within certain time period, the user will be asked to provide the range of date that is desired. For example:

```
Since year: 2019
Since month: 12
Since day: 12
Until year: 2019
Until month: 12
Until day: 31
```

Followed by a prompt asking if user wants to download posts based on hashtag(s) or based on username(s).

```
(1) Download posts based on hashtag(s)
(2) Download posts from user(s)
Enter (1) or (2):
```

After that, the user will be asked if he/she wants download only images, only videos, or both images and videos.

```
(1) Image only
(2) Video only
(3) Both Image and video
Enter (1), (2), or (3):
```

Followed by another prompt asking if user wants to download with one hashtag or two hashtag.

```
(1) Download with one hashtag
(2) Download with two hashtag
Enter (1) or (2):
```

User should only enter 2 if there are exactly two hashtags written in the `classes.txt` file.

The images and/or videos are named in `username_uploaddate` pattern.The `txt` caption files have the same naming pattern as the samples file name. Therefore it is easy for user to see the corresponding caption of a post. Example of a `txt` file named `lunali_1_2020-12-14_22-40-15.txt` is given below. `lunali_1` is the account that uploaded the post, while `2020-12-14_22-40-15` is the time that post was uploaded.

```
Mila the milanese
*
*
#dogs #dog #floof #brindle
```

Below is an example of how the directories look like when a user downloads one post with 'cat' hashtag and one post from a username called 'kyliejenner':

```
database
--google
--instagram
----hashtags
------cat
--------captions
----------_loversgifts__2021-01-06_20-44-32.txt
--------_loversgifts__2021-01-06_20-44-32_1.jpg
----users
------kyliejenner
--------captions
----------kyliejenner_2020-12-14_02-50-53_1.txt
--------kyliejenner_2020-12-14_02-50-53_1.jpg
--twitter
```

7 Twitter Scraper

Twitter scraper mainly scrapes tweets and download them into separate txt files. Beside that, if a tweet contains images, they will be saved as well. After the user chose Twitter and entered the number of samples to be downloaded, the user will be asked to indicate if he/she wants to download posts in certain time period.

```
Download posts in certain time period?
(1) Yes
(2) No
Enter (1) or (2):
```

If a user wants to download posts within certain time period, the user will be asked to provide the range of date that is desired. For example:

```
Since year: 2019
Since month: 12
Since day: 12
Until year: 2019
Until month: 12
Until day: 31
```

After that, the user will be asked if he/she wants to download tweets based on keywords or based on usernames:

```
(1) Download tweets based on keyword(s)
(2) Download tweets from user(s)
Enter (1) or (2):
```

The `txt` files that contain tweets are named in `tweetid_todaydate` pattern. A `txt` file contains the tweet, number of current likes, and the date that tweet was posted. The images that are downloaded are named in `tweetid_index` pattern, where index indicate the index of the image in case the tweet contains more than one image. An example of `txt` file name `1334663899572883457_2021-01-11.txt` is given below. `1334663899572883457` is the tweet id, while `2021-01-11` is the date that I downloaded this tweet.

```
A note from Billie on the \WHERE DO WE GO?" World Tour.
https://t.co/y23giu5agi
Likes: 67517
Posted at: 2020-12-04 01:00:45
```

Below is an example of how the directories look like when a user downloads one post with 'cat' keyword and one post from a username called 'billieilish':

```
database
--google
--instagram
--twitter
----keywords
------cat
--------media
--------1346923051829633028_2021-01-06.txt
----users
------billieeilish
--------media
----------1334663899572883457_0.jpg
--------1334663899572883457_2021-01-11.txt
```

Not all tweets contain an image, therefore it can be seen that `database/twitter/keywords/cat/media` is empty.