University of Groningen

Web Scraping Tool

Short Programming Project

User Guide

Version: 1.0

Author:
Keiko Angela Nicolasky

Contents

1 Downloading the Tool **2**

2 Installing Required Packages **2**

3 Entering Keywords/Classes to be Downloaded **2**

4 Choosing Web to Download Contents From Number of Contents to be Downloaded **2**

5 Google Scraper **3**

6 Instagram Scraper **3**

1 Downloading the Tool

The tool is available to download from a GitHub public 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/Classes 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/classes he/she wants to download, separated by newline. User can put a keyword that contain space, e.g 'cute cat'. 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.

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
Enter 1 or 2:
```

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 content(s) he/she wants to download from the chosen web.

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

Contents can be images, videos, or both. The entered number indicates how many contents per keyword/hashtag that will be downloaded. It is important to note that depending on the keywords, the actual numbers of downloaded contents may be less than the entered number. When downloading contents from Instagram, it may also be the case that the downloaded images or videos are more than the entered number.

For downloading contents that contain two hashtags, the entered number will be the number of contents 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 Scraper

Google scraper only downloads images from google image. After the user chose Google and entered the number of contents to be downloaded, the program will take some times 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.
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.

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 contents to be downloaded, the user will be asked to indicate if he/she wants to 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 program will create a path `/database/instagram`. Inside the instagram folder, there will be folder(s) name according to hashtag(s). User can find the downloaded contents of each keywords inside the corresponding folder. The files are named in `profile_date` pattern, where profile is the username of the account that posted the content, and date is the date and time the content was posted.
Additionally, inside each folder, there is also a caption folder, with `txt` files containing the caption of each posts. The `txt` files have the same naming pattern as the contents file name. Therefore it is easy for user to see the corresponding caption of a content.