Python Web Crawler

Ankeet Saha

June 10, 2023

0.1 Introduction

This report describes a Python web crawler developed as part of the DIC project. The web crawler takes a website URL as input and returns the links found on the website. It provides various command line arguments for customization.

0.2 Modules Used

- argparse
- requests
- urllib
- bs4
- lxml
- colorama
- urllib3

To ensure that the web crawler functions correctly in your system you can use the command `pip3 install <names of the Modules mentioned above>`

0.3 Usage

To run the web crawler, use the following command line arguments:

```
finalProject.py -u <url> -t <threshold> -o <output file> -H
```

The `-t` and `-o` arguments are optional, while `-H` is an optional flag.
-u is a compulsory argument

0.3.1 Command Line Arguments/Flags

- `-t`: Specifies the recursive depth up to which the website will be crawled. If not provided, the crawler will crawl until there are no more links.
- `-o`: Specifies the output file where the results will be saved. If not provided, the contents are printed to the terminal.
- `-H`: A flag that enables error handling from the server side.If not provided, all links found are printed, regardless of encountering errors or not. It handles the request exception errors.
- `-u`: This is a compulsory argument which is used to provide a url to crawl.

0.4 Crawler Characteristics

The Python web crawler developed for this project has the following characteristics:

1. Takes a website URL as input and returns the links found on the website.
2. Allows specifying the recursive depth for crawling using the `-t` command line argument.
3. Supports saving the results to an output file using the `-o` command line argument.
4. Provides error handling from the client side using the `-H` flag.

0.5 Output Format

The web crawler generates the following output:

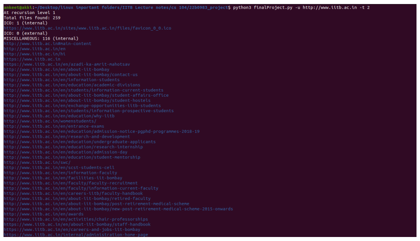
- File names in each recursive depth are printed.
- The number of files printed in each recursive depth is displayed.
- The files are segregated based on file types.
- The number of files for each file type in each recursive depth is shown.
- Each file type is further categorized as internal or external.

0.6 Customizations

The web crawler includes the following customizations:

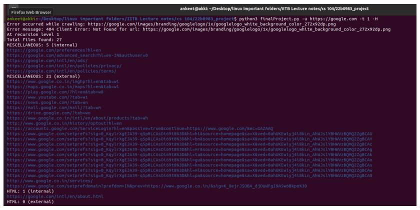
1. Valid links are colorized in blue to differentiate them from other information.
2. Files are categorized as internal or external.
3. Error handling is implemented for the following cases:
   (a) Timeout error
   (b) SSL certificate verification error
   (c) Request exception error from the client side

0.7 Screenshots



```
ankeet@akki:~/Desktop/linux important folders/IITB Lecture notes/cs 104/22b0983_project$ python3 finalProject.py -u http://www.iitb.ac.in -t 2
At recursion level 1
Total files found: 259
ICO: 1 (internal)
https://www.iitb.ac.in/sites/www.iitb.ac.in/files/favicon_0_0.ico
ICO: 0 (external)
MISCELLANEOUS: 116 (internal)
http://www.iitb.ac.in#main-content
http://www.iitb.ac.in/en
http://www.iitb.ac.in/hi
https://www.iitb.ac.in
https://www.iitb.ac.in/en/azadi-ka-amrit-mahotsav
http://www.iitb.ac.in/en/about-iit-bombay
http://www.iitb.ac.in/en/about-iit-bombay/contact-us
http://www.iitb.ac.in/en/information-students
http://www.iitb.ac.in/en/education/academic-divisions
http://www.iitb.ac.in/en/students/information-current-students
http://www.iitb.ac.in/en/about-iit-bombay/student-affairs-office
http://www.iitb.ac.in/en/about-iit-bombay/student-hostels
http://www.iitb.ac.in/en/exchange-opportunities-iitb-students
http://www.iitb.ac.in/en/students/information-prospective-students
http://www.iitb.ac.in/en/education/why-iitb
http://www.iitb.ac.in/womenstudents/
http://www.iitb.ac.in/en/entrance-exams
http://www.iitb.ac.in/en/education/admission-notice-pgphd-programmes-2018-19
http://www.iitb.ac.in/en/research-and-development
http://www.iitb.ac.in/en/education/undergraduate-applicants
http://www.iitb.ac.in/en/education/research-internship
http://www.iitb.ac.in/en/education/admission-day
http://www.iitb.ac.in/en/education/student-mentorship
http://www.iitb.ac.in/swc/
http://www.iitb.ac.in/en/scst-students-cell
http://www.iitb.ac.in/en/information-faculty
http://www.iitb.ac.in/en/facilities-iit-bombay
http://www.iitb.ac.in/en/faculty/faculty-recruitment
http://www.iitb.ac.in/en/faculty/information-current-faculty
http://www.iitb.ac.in/en/careers-iitb/faculty-handbook
http://www.iitb.ac.in/en/about-iit-bombay/retired-faculty
http://www.iitb.ac.in/en/about-iit-bombay/post-retirement-medical-scheme
http://www.iitb.ac.in/en/about-iit-bombay/new-post-retirement-medical-scheme-2015-onwards
http://www.iitb.ac.in/en/awards
http://www.iitb.ac.in/en/activities/chair-professorships
https://www.iitb.ac.in/en/about-iit-bombay/staff-handbook
https://www.iitb.ac.in/en/careers-and-jobs-iit-bombay
https://www.iitb.ac.in/internal/administration-home-page
```

(a) Example 1


Firefox Web Browser
ankeet@akki: ~/Desktop/linux important folders/IITB Lecture notes/cs 104/22b0983_project

```
ankeet@akki:~/Desktop/linux important folders/IITB Lecture notes/cs 104/22b0983_project$ python3 finalProject.py -u https://google.com -t 1 -H
Error occurred while crawling: https://google.com/images/branding/googlelogo/1x/googlelogo_white_background_color_272x92dp.png
Error message: 404 Client Error: Not Found for url: https://google.com/images/branding/googlelogo/1x/googlelogo_white_background_color_272x92dp.png
At recursion level 1
Total files found: 27
MISCELLANEOUS: 5 (internal)
https://google.com/preferences?hl=en
https://google.com/advanced_search?hl=en-IN&authuser=0
https://google.com/intl/en/ads/
https://google.com/intl/en/policies/privacy/
https://google.com/intl/en/policies/terms/
MISCELLANEOUS: 21 (external)
https://www.google.co.in/imghp?hl=en&tab=wi
https://maps.google.co.in/maps?hl=en&tab=wl
https://play.google.com/?hl=en&tab=w8
https://www.youtube.com/?tab=w1
https://news.google.com/?tab=wn
https://mail.google.com/mail/?tab=wm
https://drive.google.com/?tab=wo
https://www.google.co.in/intl/en/about/products?tab=wh
http://www.google.co.in/history/optout?hl=en
https://accounts.google.com/ServiceLogin?hl=en&passive=true&continue=https://www.google.com/&ec=GAZAAQ
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=hi&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAU
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=bn&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAY
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=te&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAc
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=mr&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAg
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=ta&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAk
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=gu&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAo
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=kn&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAs
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=ml&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCAw
https://www.google.com/setprefs?sig=0_KqylrXgEJA39-q5pRLCAsOi69t8%3D&hl=pa&source=homepage&sa=X&ved=0ahUKEwiyj4i8kLn_AhWJslYBHWVzBQMQ2ZgBCA0
http://www.google.co.in/services/
https://www.google.com/setprefdomain?prefdom=IN&prev=https://www.google.co.in/&sig=K_8ejrJSOBA_djOuWFgI9ASw6Bkpo%3D
HTML: 1 (internal)
https://google.com/intl/en/about.html
HTML: 0 (external)
```

(b) Example 2

Bibliography

[1] URL: https://www.geeksforgeeks.org/python-program-to-recursively-scrape-all-the-urls-of-

[2] URL: https://stackoverflow.com/questions/50270232/scrape-all-of-sublinks-of-a-website-recursively-in-python-using-beautiful-soup.