Web Crawler
Project CS104

Rijul Bhat

Contents

1 Introduction

Welcome to the documentation for the Python Web Crawler project.
Web crawler is a computer program that is used to search and automatically index website content and other information over the internet. The Web Crawler implemented by me is a python code which searches for links and files in the **src** and **href** attributes present in the source code of the web-page and then recursively crawls the links present in the source code in accordance to the needs of user. The program is expected to be run using Terminal or Command Prompt.

2 Overview

The libraries used in the python code of the web crawler are:
BeautifulSoup; requests; sys; os; argparse; urllib; urllib3
The output file generated(or output on terminal) after crawling has the following format:

.
.
.
At recursion level i
Total files found : $< files >$
Internal Links : $< no\_internal\_links >$
External Links : $< no\_external\_links >$
$< ext >$ files:
$< link >$
$< link >$
.
.
.
$< no\_of\_files >$ files found
.
.
.

3 Logic

The web crawler implemented follows the below logical scheme for printing the webpage found for both finite and infinite level recursion:

- If at the ith recursion level, a webpage is being referred by multiple webpages then that webpage will be printed only once at that particular recursion level. Also, if a webpage is found at ith recursion level and then even if it is found again at jth recursion level where $j > i$ it will not be crawled again. However, the webpage will be printed again so as to indicate that it was also obtained at that recursion level.
- Further details about code functioning can be found in the comments in the source code.

4 Customization

- Downloading File:
  Files of particular extensions can be downloaded. The extensions are user specified and

have to be space separated if multiple extension files are required to be downloaded. If no arguments are specified then all files(including web-pages) are downloaded. The files will be downloaded in a directory wise fashion. For instance if a website A is being crawled and we are downloading pdf files, if the pdf file say B.pdf has path A/x/y/B.pdf then B.pdf will be downloaded with path ./x/y/B.pdf on the local machine.

- Multiple Site Crawling:
  Multiple websites can be crawled at once. The specified threshold level will be applicable for all urls.
- Displaying File Size:
  File size is displayed along with the list of visited urls.
- Sorting Files with respect to File Size:
  Sorting list of displayed files with respect to file size. This will not display the file size but only sort with respect to size. Specify -s along with -x in order to get file size as well.
- Counting the internal and external links for each recursive level

5 Syntax and Documentation

- Usage

  python3 web_crawler.py [-h] -u URL [URL ...] [-t THRESHOLD] [-o OUTPUT [OUTPUT ...]] [-d [DOWNLOAD ...]] [-s] [-x]
- Options:

Options	**Description**
-h, –help	show this help message and exit
-u URL [URL ...]	provide url(s)
-t THRESHOLD	provide output file(s) for storing the result
-d [DOWNLOAD ...]	provide extension of file which you want to download or do not specify any arguments to download every file directory wise
-s, –size	provide flag if file size is also required in the output
-x, –sort	provide flag for sorting with respect to file size without getting file size in the output

6 Source Code

The source code of the web crawler can be found in this GitHub Repo.