Python Web Crawler

TEC NARAYAN BRAHMACHARI

June 14, 2023

Introduction

This report presents a Python web crawler developed as part of the DIC project. The purpose of this web crawler is to retrieve and analyze the links available on a given website. By providing a website URL as input, the crawler navigates through the web pages and extracts the links it encounters.

The Python web crawler offers several command line arguments to customize its behavior and output. These arguments allow users to specify the recursive depth of the crawling process, define an output file for saving the results, enable progress updates, and control the downloading of internal images. Additionally, the crawler can display the size of each link and segregate the links based on their extensions.

With its features and customizations, this web crawler provides a flexible and efficient solution for exploring website structures and collecting data. It can be applied in various domains, such as web scraping, link analysis, and content aggregation.

0.1 Usage

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

```
python3 web-crawler.py -u link -t 1 -d "image" -c 1 -o file_name
```

0.1.1 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.
- `-c`: If set to 1, the current variable is enabled, providing updates on the progress of the crawler.
- `-u`: This is a compulsory argument used to provide a URL to crawl.
- `-s`: Prints the link with its size printed before it.
- `-d`: Downloads all internal images. Pass the argument 'image' to activate this feature.

0.2 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. Downloads all internal images.
5. Shows the size of all links.
6. Displays progress while crawling.

0.3 Output Format

The web crawler generates the following output:

- Shows the current number of links in the record.
- Displays all internal files first, followed by external files, with their respective sizes.
- Shows what all got downloaded.
- Segregates links according to their extensions.

0.4 Customizations

The web crawler includes the following customizations:

1. Able to download all images from the given links. Helpful for data collection in terms of images.
2. Files are categorized as internal or external.
3. Shows the size related to each link.
4. Displays progress while crawling.

0.5 Bibliography

1. URL: https://www.geeksforgeeks.org/python-program-to-recursively-scrape-all-the-urls
2. URL: https://stackoverflow.com/questions/50270232/scrape-all-of-sublinks-of-a-website-recursively-in-python-using-beautiful-soup