POLYTECHNIQUE MONTRÉAL

4POINT0 PARTENARIAT POUR L'ORGANISATION DE L'INNOVATION ET DES NOUVELLES TECHNOLOGIES

How to use the Wayback_Machine_Crawler

The 4POINT0 Wayback Machine Crawler is a program that allows to get data from Wayback Machine – Internet Archive and store them in a MongoDB database. To crawl Wayback Machine is a little bit tricky, because their servers are protected if too many requests are executed in a short range of time. The 4POINT0 Wayback Machine Crawler allows to perform a gentle multi-retry crawling process in some easy steps, in order to get the maximum number of pages in the shorter time frame as possible (more or less 10 000 pages per hour).

To execute the program, you will use your terminal to execute a python script. The following code of line is the basic line to launch the program:

```
python -m Scripts.Wayback_Machine_Crawler -i Import/data_test.txt -uri mongodb://localhost -db testcoll -
coll data -coll_err data_err
```

In the previous line of code you have to modify the value of those arguments (yellow marked strings) with your personal information. The list of all possible arguments can be got with --help.

Requisites

In order to use the 4POINT0 Wayback Machine Crawler, some requisites are necessary.

1. Install MongoDB on your machine.
2. Install Python on your machine (we recommend Anaconda distribution)
3. Install all the following modules on your Python distribution. The following list could be not complete:
   a) pymongo
   b) tqdm
   c) fake_useragent
   d) json
   e) BeautifulSoup
4. Have a .txt file with your list of websites to crawl.
   a) The .txt file have to contain the website domain **without** the http(s)://www. prefix and the year for which the information will be crawled. These two inputs have to be separated by a comma-space (“, ”).
   b) Here an example:

```
1 caprock.com, 2007
2 coherencecollaborative.com, 2007
3 culinarycollective.com, 2007
4 dansko.com, 2007
```

5. dddd

Documentation

To execute the program you will send a command from your terminal. Here the steps which has to be done:

1. Download the code from the GitHub repository. Then, unzip it if necessary.
2. Open your terminal and navigate on the main directory of the Wayback_Machine_Crawler. You should have the folders Scripts, Import and LOG as children of this directory.
3. Start your version of the MongoDB service on your system (for example: sudo systemctl start mongod)
4. Execute the following line code to get help information about arguments and usage of the crawler:

```
python -m Scripts.Wayback_Machine_Crawler --help
```

Details of the crawler

The 4POINT0 Wayback_Machine_Crawler begins to crawl from the input .txt file. Then, it try to solve some error that occurred in a two step process. The first one use the main crawler for those errors occurred during the first access to a domain. The second process try to solve errors occurred for each url.

Davide Pulizzotto