Homework Four

Theory of Computation

2015

Instructor: Kun-Ta Chuang, email: ktchuang@mail.ncku.edu.tw.

Important Note:

Please remember to upload your homework to server before 6/18 (Thursday) 11:59 p.m., and the server information will be announced later. You are not allowed to revise or submit the homework after 6/18 (Thursday) 11:59 p.m., the server will be closed then.

Please upload your python program complying with these rules:

1. Please put the main function in "tocHw4.py", and this homework is only allowed to be written in python.
2. Please put "all" your python source file at the "hw4" directory (you should create the directory at first).
3. You can find the testing data in /home/toc/toc4/input.txt.
4. You program must accept two argument, which are "input_file_name, query". If no input is given, please show messages and reject the execution.
5. Write your basic information in the start of the python source code, including your name, your student number and brief description of your code.

Homework 4: Given a Json file containing 100 pieces of data about the web page information and the specific query, please scan the whole data and find the web pages containing the query name. For these web pages, select the out-links excluding the query name and compute the types of filename extension of these web pages and the number of each type. Finally print out the types of filename extension of these web pages and the number of each type.

Arguments:

1. Input_file_name: Specify the testing data name in /home/toc/toc4/. For example: "/home/toc/toc4/input.txt"
2. web_site: Specify query name. For example: 0xxx.in.

Running Examples:

Input: python tocHw4.py /home/toc/toc4/input.txt 0xxx.in

Output:

```
html:1
jpg:1
```

Note: 1. Some web pages may have no out-links.

2. One row of json format is one web page in our input file.

3. If no web pages matching the given query, please show the "Page not found!" message. On the other hand, if there are no types of filename extension of these web pages not

containing the query name, please show the “Type not found!” message.

4. Only the out-links in array named ”Links” are the ones you should process.
5. The key name of a out-link could be "href" or "url". And different from hw3, only the url of out-link containing the “http://” or “https://” is considered in this homework.
6. To identify the filename extension, there exists the case of url having parameters and need to be parsed the filename extension before “?”.
7. If you want to practice the python coding of final project, please apply the python 2.7 to write your homework. After all, we only and strongly suggest the python 2.7 in the final project.

8.The top ten: the students who write the program with the correct answer and lowest execution time can obtain extra bonus.

Input File Description: (The important value of each web page in the input file)

Suppose the given query name is “news.bbc.co.uk”.

1. The URL of this web page:

"WARC-Target-URI" : http://news.bbc.co.uk/2/hi/africa/3414345.stm

2. The URL of the out-links of this web page:

```
"Links" : [     {
                     "href" : "/css/screen/shared/styles.css",
                     "path" : "STYLE/#text"
                },  // skip; no https or http
                {
                     "title" : "At least 66 dead after another powerful earthquake hits Nepal",
                     "path" : "A@/href",
                     "url" : "http://edition.cnn.com/nepal-earthquake/index.html"
                },  // filename extension is html
                {
                     "title" : "At least 66 dead after another powerful earthquake hits Nepal",
                     "path" : "A@/href",
                     "url" : "http://www.cnn.com /smw/index.php?title=Special"
                },  // filename extension is php
                {
                     "title" : "At least 66 dead after another powerful earthquake hits Nepal",
                     "path" : "A@/href",
                     "url" : "http://wp-content/ news.bbc.co.uk/index.php?id=1"
                } //skip; contain the specific query name
           ],
```

Take this web page as an example, there are two types of filename extension “html” and “php” with both 1 counting value.

Finally:

When you submit your code to the server, please make sure if your code can execute in the linux environment. And also, both of the “re” and “json” library are available in our python environment, therefore, no external system library can be used in this homework.

Good luck!