Homework Three

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 "tocHw3.py", this homework is only allowed to be written in python.
2. Please put "all" your python source file at the "hw3" directory (you should create the directory at first).
3. You can find the testing data in /home/toc/toc3/input.txt.
4. You program must accept two argument, which are "input_file_name, top _k". 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 3: Given a Json file containing 100 pieces of data about the web page information, please scan the whole data and choose the top k web pages with the highest number of out-links. Finally print out these top k web pages and the number of their out-links.

Arguments:

1. Input_file_name: Specify the testing data name in /home/toc/toc3/. For example: "/home/toc/toc3/input.txt"
2. k: Specify the number of top k. For example: 3.

Running Examples:

Input: python tocHw3.py /home/toc/toc3/input.txt 3

Output:

http://6dollarshirts.com/news/2012/09/25/slogans-ideas-t-shirt-contest/product.php?productid=11496:927

http://6dollarshirts.com/news/2012/10/12/spread-breast-cancer-awarenessfree-shades/product.php?productid=11951:927

http://6dollarshirts.com/news/2010/08/09/6dollarshirts-com-6-year-anniversary-party/steve_pinata_b/product.php?productid=12135:927

http://6dollarshirts.com/news/mat/product.php?productid=11538:927

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. Only the out-links in array ”Links” are the ones you should process.
4. The key named of a out-link could be "href" or "url".
5. If the out-link amounts of some web pages are the same, the order of these pages is not considered. And the web pages with the same number of out-links as the top k should be also output as the answer, even if the length of answer is over k.
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)

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"
                     },
                     {
                          "title" : "Kifaransa kwa Afrika",
                          "path" : "A@/href",
                          "url" : "http://www.bbc.co.uk/french"
                     }
                     {
                     }
                ],
```

Take this web page as an example, the number of out-links is 2.

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!