Introduction to Computer Networks

Lab 1

1. Description

Write a program which can show all the hyperlinks in a given web page. The main objective is to practice socket programming and try to use the HTTP protocol.

2. Requirements

(a) Input: The URL of the desired webpage without "http://". e.g. For http://can.cs.nthu.edu.tw/index.php, you only need to type can.cs.nthu.edu.tw/index.php. Note that http://can.cs.nthu.edu.tw and http://can.cs.nthu.edu.tw/LAB/ are also available URLs, while you don't need to handle these formats in this lab.

(b) Output: Print all hyperlinks and the number of hyperlinks in the given webpage. Note that only <a href="xxx"> should be counted, whereas <link href="xxx"> shouldn't.

(c) The code you write will be compiled and tested under the given vmware environment (Ubuntu 20.04) . Test the code on the machine by yourself before final submission.

3. Example

```
canlab@canlab-All-Series:~/Desktop$ ./lab1
Enter the hostname: can.cs.nthu.edu.tw/index.php
socket: Start send HTTP request
socket: Start read the response
socket: Finish read to buffer
======== Hyperlinks ========
index.php
members.php
LAB/
gallery.php
contact.php
http://web.cs.nthu.edu.tw/files/14-1015-143485,r109-1.php?Lang=zh-tw
http://www.nthu.edu.tw
http://web.cs.nthu.edu.tw/bin/home.php
http://www.com.nthu.edu.tw/
http://www.highimpact-seo.co.uk/
```

There are 10 hyperlinks in http://can.cs.nthu.edu.tw/.

4. Hint:

(a) Refer to "Linux socket tutorial" for information about socket programming on Linux.

(b) Connect to the web server whose host name is given.

- Convert the host name (e.g., can.cs.nthu.edu.tw) into IP address.
- Set the protocol. (e.g., always use TCP protocol for HTTP)

(c) Get the HTML source code of the web page.

- Send a HTTP GET request message to the server. (GET xxx …\r\n\r\n)
- Receive the response from the server.

(d) Show all hyperlinks in the web page.

- <a href="xxxx">
- If the string starts with "href" and the previous token ends with "a", then you find a hyperlink.

5. Submission

(a) Please provide a pdf file to show what functionalities your homework has.

- For example, is it able to be compiled by gcc? Does it meet all requirements?
- If you can run your C program, please provide a screenshot to show how it works just like the examples in this document.

(b) Compress the C source file(s) and related files (including readme.pdf) into 學號_作業.zip (ex: 111012345_lab1.zip).

(c) Upload your zip file to eeclass.

(d) Discussion is encouraged. However, plagiarism is not allowed. We will use, e.g., "Moss" for similarity comparison and 0 points will be given if plagiarism.

(e) You should submit your assignment by the deadline, or your assignment will not be graded, meaning that you will receive zero points.