浙江大学

本科实验报告

课程名称：	计算机网络基础
实验名称：	实现一个轻量级的 WEB 服务器
姓　　名：	董展辰
学　　院：	公共管理学院
专　　业：	信息资源管理
学　　号：	3200104203
指导教师：	陆系群

2022 年　　12 月　　29 日

浙江大学实验报告

实验名称：实现一个轻量级的 WEB 服务器　实验类型：编程实验

同组学生：董展辰　实验地点：计算机网络实验室

一、 实验目的

深入掌握 HTTP 协议规范，学习如何编写标准的互联网应用服务器。

二、 实验内容

- 服务程序能够正确解析 HTTP 协议，并传回所需的网页文件和图片文件
- 使用标准的浏览器，如 IE、Chrome 或者 Safari，输入服务程序的 URL 后，能够正常显示服务器上的网页文件和图片
- 服务端程序界面不做要求，使用命令行或最简单的窗体即可
- 功能要求如下：
  1. 服务程序运行后监听在 80 端口或者指定端口
  2. 接受浏览器的 TCP 连接（支持多个浏览器同时连接）
  3. 读取浏览器发送的数据，解析 HTTP 请求头部，找到感兴趣的部分
  4. 根据 HTTP 头部请求的文件路径，打开并读取服务器磁盘上的文件，以 HTTP 响应格式传回浏览器。要求按照文本、图片文件传送不同的 Content-Type，以便让浏览器能够正常显示。
  5. 分别使用单个纯文本、只包含文字的 HTML 文件、包含文字和图片的 HTML 文件进行测试，浏览器均能正常显示。
- 本实验可以在前一个 Socket 编程实验的基础上继续，也可以使用第三方封装好的 TCP 类进行网络数据的收发
- 本实验要求不使用任何封装 HTTP 接口的类库或组件，也不使用任何服务端脚本程序如 JSP、ASPX、PHP 等

三、 主要仪器设备

联网的 PC 机、Wireshark 软件、Visual Studio、gcc 或 Java 集成开发环境。

四、 操作方法与实验步骤

- 阅读 HTTP 协议相关标准文档，详细了解 HTTP 协议标准的细节，有必要的话使用 Wireshark 抓包，研究浏览器和 WEB 服务器之间的交互过程
- 创建一个文档目录，与服务器程序运行路径分开
- 准备一个纯文本文件，命名为 test.txt，存放在 txt 子目录下面
- 准备好一个图片文件，命名为 logo.jpg，放在 img 子目录下面
- 写一个 HTML 文件，命名为 test.html，放在 html 子目录下面，主要内容为：

```
<html>
    <head><title>Test</title></head>
    <body>
        <h1>This is a test</h1>
        <img src="img/logo.jpg">
        <form action="dopost" method="POST">
            Login:<input name="login">
            Pass:<input name="pass">
            <input type="submit" value="login">
        </form>
    </body>
</html>
```

- 将 test.html 复制为 noimg.html，并删除其中包含 img 的这一行。
- 服务端编写步骤（**需要采用多线程模式**）

a) 运行初始化，打开 Socket，监听在指定端口（**请使用学号的后 4 位作为服务器的监听端口**）

b) 主线程是一个循环，主要做的工作是等待客户端连接，如果有客户端连接成功，为该客户端创建处理子线程。该子线程的主要处理步骤是：

1. 不断读取客户端发送过来的字节，并检查其中是否连续出现了 2 个回车换行符，如果未出现，继续接收；如果出现，按照 HTTP 格式解析第 1 行，分离出方法、文件和路径名，其他头部字段根据需要读取。

✧ **如果解析出来的方法是 GET**

2. 根据解析出来的文件和路径名，读取响应的磁盘文件（该路径和服务器程序可能不在同一个目录下，需要转换成绝对路径）。如果文件不存在，第 3 步的响应消息的状态设置为 404，并且跳过第 5 步。
3. 准备好一个足够大的缓冲区，按照 HTTP 响应消息的格式先填入第 1 行（状态码=200），加上回车换行符。然后模仿 Wireshark 抓取的 HTTP 消息，填入必要的几行头部（需要哪些头部，请试验），其中不能缺少的 2 个头部是 Content-Type 和 Content-Length。Content-Type 的值要和文件类型相匹配（请通过抓包确定应该填什么），Content-Length 的值填写文件的字节大小。
4. 在头部行填完后，再填入 2 个回车换行
5. 将文件内容按顺序填入到缓冲区后面部分。

✧ **如果解析出来的方法是 POST**

6. 检查解析出来的文件和路径名，如果不是 dopost，则设置响应消息的状态为 404，然后跳到第 9 步。如果是 dopost，则设置响应消息的状态为 200，并继续下一步。
7. 读取 2 个回车换行后面的体部内容（长度根据头部的 Content-Length 字段的指示），并提取出登录名（login）和密码（pass）的值。如果登录名是你的学号，密码是学号的后 4 位，则将响应消息设置为登录成功，否则将响应消息设置为登录失败。
8. 将响应消息封装成 html 格式，如

<html><body>响应消息内容</body></html>

9. 准备好一个足够大的缓冲区，按照 HTTP 响应消息的格式先填入第 1 行（根据前面的情况设置好状态码），加上回车换行符。然后填入必要的几行头部，其中不能缺少的 2 个头部是 Content-Type 和 Content-Length。Content-Type 的值设置为 text/html，如果状态码=200，则 Content-Length 的值填写响应消息的字节大小，并将响应消息填入缓冲区的后面部分，否则填写为 0。
10. 最后一次性将缓冲区内的字节发送给客户端。
11. 发送完毕后，关闭 socket，退出子线程。

c) 主线程还负责检测退出指令（如用户按退出键或者收到退出信号），检测到后即通知并等待各子线程退出。最后关闭 Socket，主程序退出。

- 编程结束后，将服务器部署在一台机器上（本机也可以）。在服务器上分别放置纯文本文件（.txt）、只包含文字的测试 HTML 文件（将测试 HTML 文件中的包含 img 那一行去掉）、包含文字和图片的测试 HTML 文件（以及图片文件）各一个。
- 确定好各个文件的 URL 地址，然后使用浏览器访问这些 URL 地址，如 http://x.x.x.x:port/dir/a.html，其中 port 是服务器的监听端口，dir 是提供给外部访问的路径，请设置为与文件实际存放路径不同，通过服务器内部映射转换。
- 检查浏览器是否正常显示页面，如果有问题，查找原因，并修改，直至满足要求
- 使用多个浏览器同时访问这些 URL 地址，检查并发性

五、 实验数据记录和处理

请将以下内容和本实验报告一起打包成一个压缩文件上传：

- **源代码：需要说明编译环境和编译方法，如果不能编译成功，将影响评分**
- **可执行文件：可运行的.exe 文件或 Linux 可执行文件**

以下实验记录均需结合屏幕截图（截取源代码或运行结果），进行文字标注（看完请删除本句）。

- 服务器的主线程循环关键代码截图（解释总体处理逻辑，省略细节部分）

```
47      while(1)
48      {
49          epoll_event events[MAX_EVENT];
50          int n = epoll::get_instance() -> wait(events);
51          printf("looping...\n");
52          for(int i = 0; i < n; i++)
53          {
54              epoll_event ev = events[i];
55              //printf("fd = %d\n",ev.data.fd);
56              if(ev.data.fd == serv_sock)···
73              else if(ev.events & EPOLLIN)···
84              else if(ev.events & EPOLLOUT)···
```

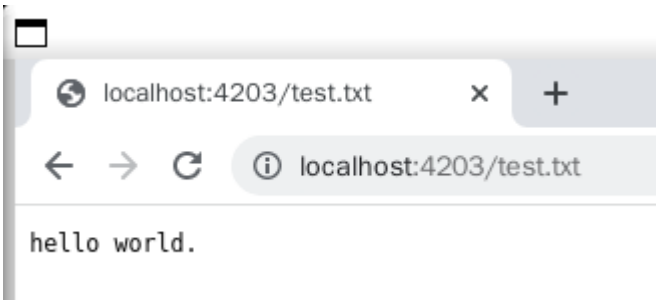
- 服务器的客户端处理子线程关键代码截图（解释总体处理逻辑，省略细节部分）

```
62  void http::process()
63  {
64
65      /* Organize answer package according to request type. */
66      char request_type[5];
67      strncpy(request_type,package,3);
68      request_type[3] = '\0';
69      if(strcmp(request_type,"GET") == 0){···
86      else{
87          strncpy(request_type,package,4);
88          request_type[4] = '\0';
89          if(strcmp(request_type,"POST") == 0){
90              char request_path[128];
91              int idx = 5;
92              while(package[idx]!=' '){
93                  request_path[idx-5] = package[idx];
94                  idx++;
95              }
96              request_path[idx-5] = '\0';
97
98              smatch m;
99              regex e("\r\n\r\n");
100             regex_search(package_str, m, e);
101             printf("The posted info is: %s\n",m.suffix().str().c_str());
102             if(strcmp(request_path,"/dopost") == 0){···
118             else{
119                 char filepath[128] = "./root/inforecved.html";
120                 response(filepath);
121             }
122             state = 1;
123         }
124         else{
125             state = 2;
126         }
127     }
128 }
```

- 服务器运行后，用 netstat –an 显示服务器的监听端口

```
[dzc@MSI MyWebServer]$ ./main.o &
[1] 3718
the serv_sock is: 3

Listening...
[dzc@MSI MyWebServer]$ netstat -an | grep :4203
tcp        0      0 127.0.0.1:4203          0.0.0.0:*               LISTEN
```

- 浏览器访问纯文本文件（.txt）时，浏览器的 URL 地址和显示内容截图。



服务器上文件实际存放的路径：

./root/test.txt

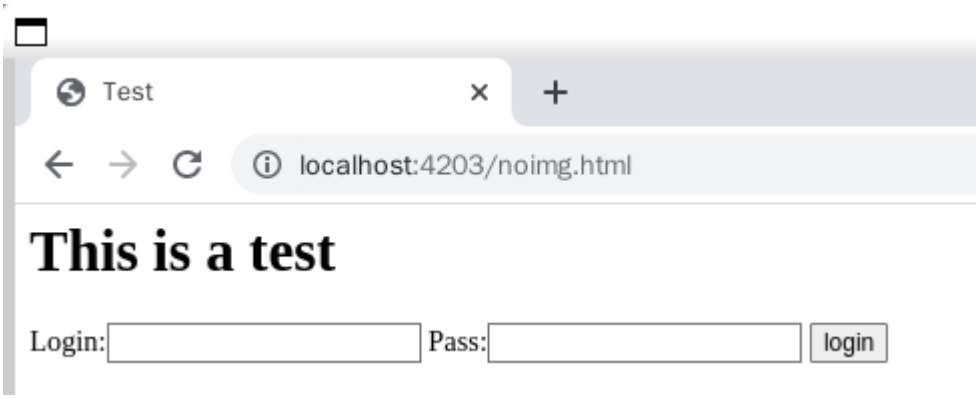
服务器的相关代码片段：

```
if(strcmp(request_type,"GET") == 0){

    char request_path[128];
    int idx = 4;
    while(package[idx]!=' '){
        request_path[idx-4] = package[idx];
        idx++;
    }
    request_path[idx-4] = '\0';
    if(strcmp(request_path,"/") == 0) strcpy(request_path,"/index.html");
    //mmap_file file("./root/index.html");
    char filepath[128] = "./root";
    strcat(filepath,request_path);

    response(filepath);
    state = 1;
}
```

Wireshark 抓取的数据包截图（通过跟踪 TCP 流，只截取 HTTP 协议部分）：

实验平台为 Windows Subsystem for Linux（WSL2），无法使用 Wireshark 抓包。

- 浏览器访问只包含文本的 HTML 文件时，浏览器的 URL 地址和显示内容截图。



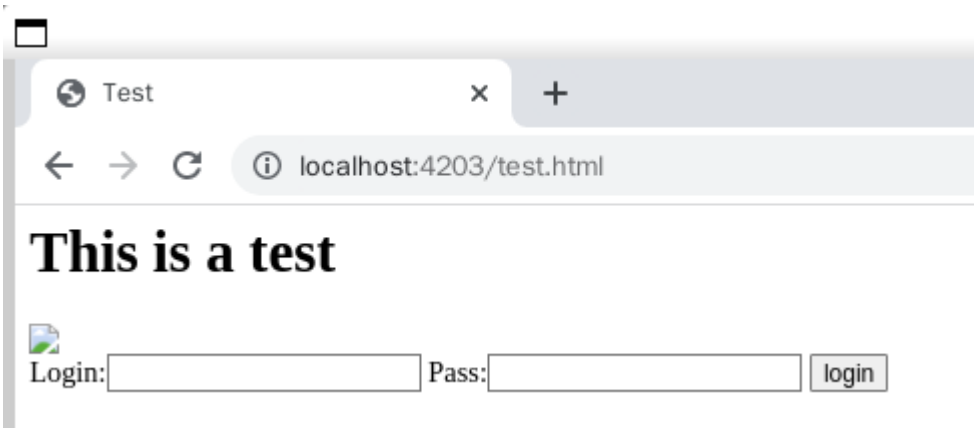
服务器文件实际存放的路径：

./root/noimg.html

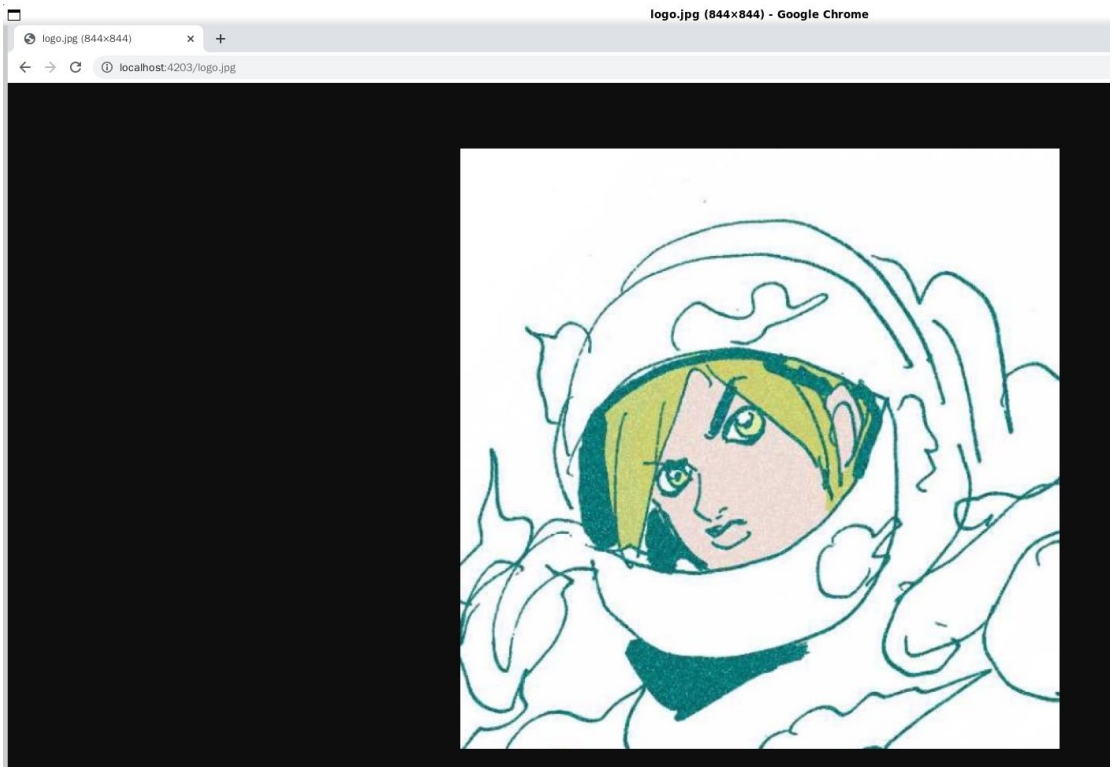
Wireshark 抓取的数据包截图（只截取 HTTP 协议部分，包括 HTML 内容）：

实验平台为 Windows Subsystem for Linux（WSL2），无法使用 Wireshark 抓包。

- 浏览器访问包含文本、图片的 HTML 文件时，浏览器的 URL 地址和显示内容截图。



（单独请求图片可以显示，但嵌入 HTML 后无法显示，原因正在寻找。）



服务器上文件实际存放的路径：

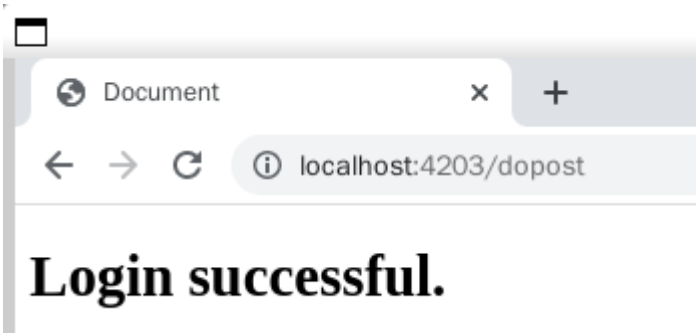
./root/test.html

./root/logo.jpg

Wireshark 抓取的数据包截图（只截取 HTTP 协议部分，包括 HTML、图片文件的部分内容）：

实验平台为 Windows Subsystem for Linux（WSL2），无法使用 Wireshark 抓包。

- 浏览器输入正确的登录名或密码，点击登录按钮（login）后的显示截图。



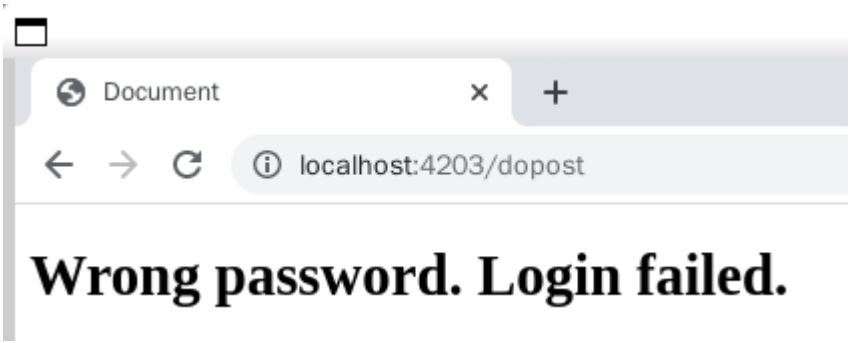
服务器相关处理代码片段：

```
if(strcmp(request_type,"POST") == 0){
    char request_path[128];
    int idx = 5;
    while(package[idx]!=' '){
        request_path[idx-5] = package[idx];
        idx++;
    }
    request_path[idx-5] = '\0';

    smatch m;
    regex e("\r\n\r\n");
    regex_search(package_str, m, e);
    printf("The posted info is: %s\n",m.suffix().str().c_str());
    if(strcmp(request_path,"/dopost") == 0){
        string posted_info = m.suffix().str();
        string login, pass;
        int idx = 6; // Cutoff "login="
        while(posted_info[idx] != '&'){
            login.push_back(posted_info[idx]);
            idx++;
        }
        pass = posted_info.substr(idx+6); // Cutoff "&pass="
        cout << "login = " << login << ", pass = " << pass << endl;

        char filepath[128] = {0};
        if(login == "3200104203" && pass == "4203") strcat(filepath,"./root/login_successful.html");
        else strcat(filepath,"./root/login_failed.html");
        response(filepath);
    }
```

Wireshark 抓取的数据包截图（HTTP 协议部分）

实验平台为 Windows Subsystem for Linux（WSL2），无法使用 Wireshark 抓包。

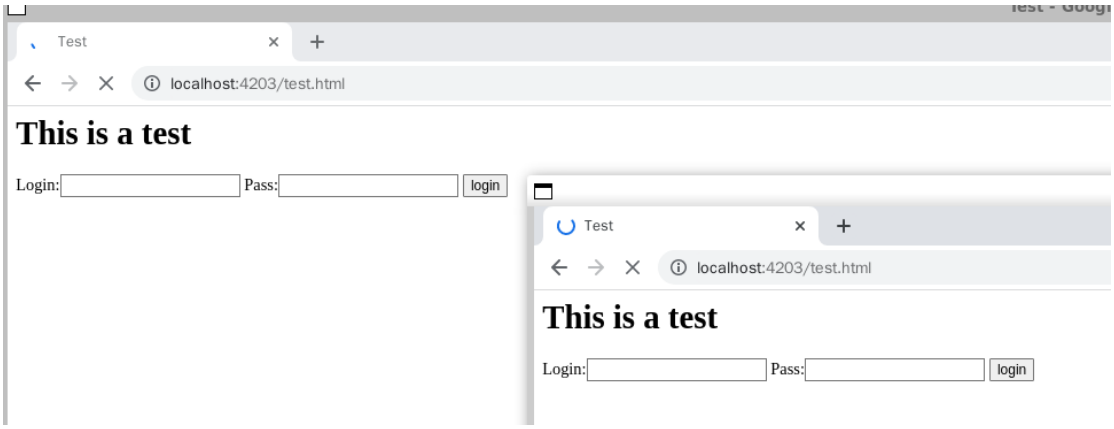
- 浏览器输入错误的登录名或密码，点击登录按钮（login）后的显示截图。



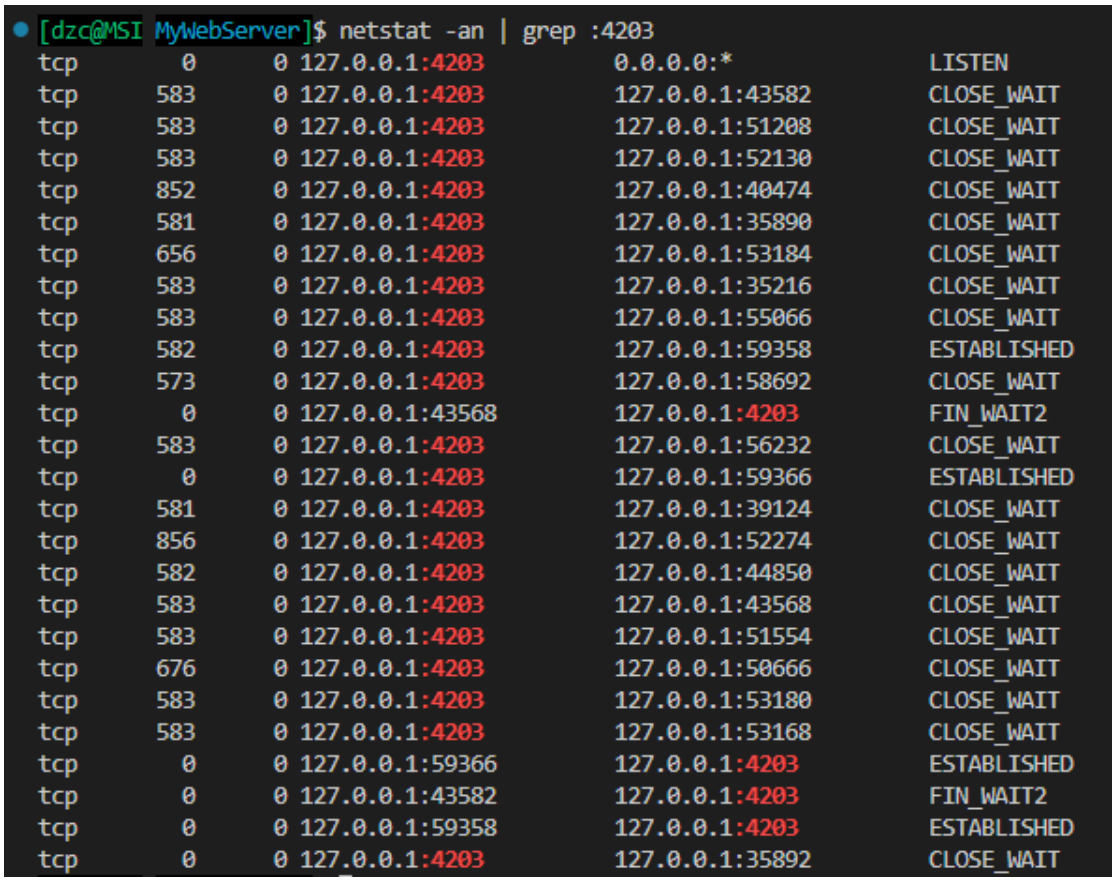
- Wireshark 抓取的数据包截图（HTTP 协议部分）

  实验平台为 Windows Subsystem for Linux（WSL2），无法使用 Wireshark 抓包。

- 多个浏览器同时访问包含图片的 HTML 文件时，浏览器的显示内容截图（将浏览器窗口缩小并列）



- 多个浏览器同时访问包含图片的 HTML 文件时，使用 netstat –an 显示服务器的 TCP 连接（截取与服务器监听端口相关的）



```
[dzc@MSI MyWebServer]$ netstat -an | grep :4203
tcp        0      0 127.0.0.1:4203          0.0.0.0:*               LISTEN
tcp      583      0 127.0.0.1:4203          127.0.0.1:43582         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:51208         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:52130         CLOSE_WAIT
tcp      852      0 127.0.0.1:4203          127.0.0.1:40474         CLOSE_WAIT
tcp      581      0 127.0.0.1:4203          127.0.0.1:35890         CLOSE_WAIT
tcp      656      0 127.0.0.1:4203          127.0.0.1:53184         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:35216         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:55066         CLOSE_WAIT
tcp      582      0 127.0.0.1:4203          127.0.0.1:59358         ESTABLISHED
tcp      573      0 127.0.0.1:4203          127.0.0.1:58692         CLOSE_WAIT
tcp        0      0 127.0.0.1:43568         127.0.0.1:4203          FIN_WAIT2
tcp      583      0 127.0.0.1:4203          127.0.0.1:56232         CLOSE_WAIT
tcp        0      0 127.0.0.1:4203          127.0.0.1:59366         ESTABLISHED
tcp      581      0 127.0.0.1:4203          127.0.0.1:39124         CLOSE_WAIT
tcp      856      0 127.0.0.1:4203          127.0.0.1:52274         CLOSE_WAIT
tcp      582      0 127.0.0.1:4203          127.0.0.1:44850         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:43568         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:51554         CLOSE_WAIT
tcp      676      0 127.0.0.1:4203          127.0.0.1:50666         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:53180         CLOSE_WAIT
tcp      583      0 127.0.0.1:4203          127.0.0.1:53168         CLOSE_WAIT
tcp        0      0 127.0.0.1:59366         127.0.0.1:4203          ESTABLISHED
tcp        0      0 127.0.0.1:43582         127.0.0.1:4203          FIN_WAIT2
tcp        0      0 127.0.0.1:59358         127.0.0.1:4203          ESTABLISHED
tcp        0      0 127.0.0.1:4203          127.0.0.1:35892         CLOSE_WAIT
```

六、 实验结果与分析

根据你编写的程序运行效果，分别解答以下问题（看完请删除本句）：

- HTTP 协议是怎样对头部和体部进行分隔的？

  头部和体部中间用一个空行分隔。（\r\n）

- 浏览器是根据文件的扩展名还是根据头部的哪个字段判断文件类型的？

  浏览器是根据头部的 Content-Type 字段判断文件类型的。

- HTTP 协议的头部是不是一定是文本格式？体部呢？

  HTTP 协议的头部一定是文本格式，体部不一定，也可以是图片等多种格式。

- POST 方法传递的数据是放在头部还是体部？两个字段是用什么符号连接起来的？

  POST 方法传递的数据是放在体部，两个字段用&连接。

七、　讨论、心得