PROJECT 3: HTTP Server

HTTP 1.1 [RFC 2616] defines the following methods: OPTIONS, GET, HEAD, POST, PUT, DELETE, TRACE and CONNECT. The goal of this project is to implement a minimal HTTP 1.1 server supporting and implementing only the GET and HEAD methods. This protocol typically runs on top of TCP on port 80.

1 Description

The HTTP protocol is a request/response protocol:

1. A client sends a request to the server in the form of a request method, URI, and protocol version, followed by a MIME-like message containing request modifiers, client information, and possibly body content over a connection with a server.
2. The server responds with a status line, including the message's protocol version and a success or error code, followed by a MIME-like message containing server information, entity metainformation, and possibly entity-body content.

Where a URI (Uniform Resource Identifier) is either a URL (Uniform Resource Location) or a URN (Uniform Resource Name). Throughout this document the following notation is used: SP is a blank space and CRLF is a carriage return followed by a line feed character. URIs have been known by many names: WWW addresses, Universal Document Identifiers, Universal Resource Identifiers , and finally the combination of Uniform Resource Locators (URL) and Names (URN). As far as HTTP is concerned, Uniform Resource Identifiers are simply formatted strings which identify--via name, location, or any other characteristic--a resource.

2 Client Request

The general form for an HTTP/1.1 request is:

Method SP Request-URI SP HTTP/1.1 CRLF
([general-header line | request-header line | entity-header line] CRLF)*
CRLF
[message body]

For this project you are only required to implement the GET and HEAD method. Note that a request may be followed by several other lines (some of them are mandatory as explained below) ended with a CRLF. The end of the header lines is indicated by the client by sending an additional CRLF. The [message body] part will not be used in this project. The message body part is where you would place data being sent to the server, e.g. if we were implementing the PUT method, the message body would contain the data to be sent to the server.

Any HTTP/1.1 compliant client must include a Host request-header field identifying the internet host (the server) they are connecting to and the port number:
Host: hostname[:port]
The port number must be included unless the server is using the default port 80. For the header lines, the notation (…)* means zero or more of the strings enclosed in parenthesis. If the machine where the server is running does not have a DNS name, you may use the IP address for hostname.
Your server must check for the Host request header field before serving the request. If there is no Host request-header, the server must return an error code as specified in the following section.

In summary, a minimal and valid GET request on a server named neo.mcs.utulsa.edu running on port 16405 for a file or resource named index.html will look like:

GET /index.html HTTP/1.1 CRLF
Host: neo.mcs.utulsa.edu:16405 CRLF
CRLF

The same request as sent by MS Internet Explorer 6.0 looks like:
GET /index.html HTTP/1.1 CRLF
Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, application/vnd.ms-excel, application/vnd.ms-powerpoint, application/msword, */* CRLF
Accept-Language: en-us CRLF
Accept-Encoding: gzip, deflate CRLF
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; Q312461) CRLF
Host: neo.mcs.utulsa.edu:16405 CRLF
Connection: Keep-Alive CRLF
CRLF

Note in this case that the client is sending lots of information to the server. It basically notifies the server what kinds of file Internet Explorer accepts (Accept lines) and a User-Agent line that notifies the server that the client is IE 6.0 running on a Windows 2000 machine.

Remember that there could be more lines following the GET method request. Your server must look among those lines for a Host request-header field before it is served (and return an error code if the line is missing). Your server can safely ignore the lines that are not strictly required by the server in order to process a client's request.

3 Server Response

After receiving and interpreting a request message, a server responds with an HTTP response message. The general form of a response message consists of several lines: (i) a status line followed by a CRLF, (ii) zero or more header lines followed by a CRLF (also called entity headers), (iii) a CRLF indicating the end of the header lines and (iv) a message body if necessary containing the requested data. The general form of the response message then is:

HTTP/1.1 SP StatusCode SP ReasonPhrase CRLF
([(general-header line | response-header line | entity-header line) CRLF) *
CRLF
[message-body]

Note: the only difference between a server response to a GET and a server response to a HEAD is that when the server responds to a HEAD request, the message body is empty (only the header lines associated with the request are sent back to the client).

3.1 Status Line

HTTP/1.1 SP StatusCode SP ReasonPhrase CRLF

The StatusCode element is a 3-digit integer result code of the attempt to understand and satisfy the request. These codes are fully defined in section 10 of RFC 2616. The ReasonPhrase is intended to give a short textual description of the StatusCode. The StatusCode is intended for use by automata and the ReasonPhrase is intended for the human user. The client is not required to examine or display the ReasonPhrase.

The first digit of the StatusCode defines the class of response. The last two digits do not have any categorization role. There are 5 values for the first digit:

- 1xx: Informational - Request received, continuing process.
- 2xx: Success - The action was successfully received, understood, and accepted.
- 3xx: Redirection - Further action must be taken in order to complete the request.
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled.
- 5xx: Server Error - The server failed to fulfill an apparently valid request.

The following table summarizes the status codes and reason phrases that your server MUST implement:

StatusCode	ReasonPhrase
200	OK
400	Bad Request
404	Not Found
501	Not Implemented

A 200 OK response indicates the request has succeeded and the information returned by the server with the response is dependent on the method used in the request. If the message is in response to a GET method, then the message body contains the data associated with the request resource. If the message is in response to a HEAD method, then only entity header lines are sent without any message body.

A 400 Bad Request message indicates that a request could not be understood by the server due to malformed syntax (the client should not repeat the request without modification).

A 404 Not Found message indicates the server has not found anything matching the requested resource (this is one of the most common responses).

A 501 Not Implemented message indicates that the server does not support the functionality required to fulfill the request. Your server should respond with this message when the request method corresponds to one of the following: OPTIONS, POST, PUT, DELETE, TRACE and CONNECT (the methods you are NOT implementing in this project).

3.2 Entity headers

There are several types of header lines. The intention of these lines is to provide information to the Client when responding to a request. Your server MUST implement (and send back to the client) at least the following three header lines:

Server: ServerName/ServerVersion
Content-Length: lengthOfResource
Content-Type: typeOfResource

The Server response-header field contains information about the software used by the origin server (the server you are implementing for this project) to handle the request. An example is:
Server: cs4333httpserver/1.0.2

The Content-Length entity-header field indicates the size of the entity-body, in decimal number of OCTETs, sent to the recipient or, in the case of the HEAD method, the size of the entity-body that would have been sent had the request been a GET. An example for lengthOfResource=3495 is:
Content-Length: 3495

The Content-Type entity-header field indicates the media type of the message-body sent to the recipient or, in the case of the HEAD method, the media type that would have been sent had the request been a GET. Your server must be able to report to the client the following media types:

File name	File Type	Content-Type
*.html, *.htm	html	text/html
*.gif	gif	image/gif
*.jpg, *.jpeg	jpeg	image/jpeg
*.pdf	pdf	application/pdf

An example for typeOfResource when responding to a GET request method for a file named index.html is:
Content-Type: text/html

As illustrated above, the end of the entity header lines is indicated by sending an additional CRLF.

3.3 Message body

The message-body (if any) of an HTTP message is used to carry the entity-body associated with the request or response (typically the file being requested). I suggest you use the FileInputStream Java class to read the contents of a requested file.

4 Implementation

Your implementation MUST follow the following guidelines:

1. Use the Java Socket API for this project.
2. You can use as many classes as you want, but the class containing the main starting point should be named HttpServer.java and it should accept a number as a command line option indicating the port number to be used by your server. For instance, a valid invocation to start the server on port 16405 is:
   java HttpServer 16405
3. Do not use packages in your implementation. I should be able to test your program by putting all your classes in one flat directory.
4. Your server should assume the existence of a directory named public_html in the directory where your class and source files are stored. Under no circumstances your server will serve any files not in the public_html directory (or any subdirectory below public_html). We do this for security reasons. If you are not careful, your server would be able to serve any requested files with a valid resource name. The public_html directory is the root directory for any files served by the server. See the Security Considerations of RFC 2616 (Section 15) for additional information. **Note:** the public_html directory only serves as the base directory for the server and should not be included when forming the URL.
5. Echo back to the screen the entire contents of any requests your server receives and any responses up to, but not including the message body (if there is one to transmit).
6. Your server should be multithreaded and must be able to handle multiple concurrent requests.
7. It is advised that you do not run your server on port 80. There are plenty of Internet worms and hackers scanning machines and looking for port 80.

5 Testing

A good way to test your server is by using any Web browser. Simply point it to the machine where you are running the server, indicate the port number and the requested file. For example, if a server is running on a machine named neo.mcs.utulsa.edu on port 16405 and we request a file named index.html, you would type the following in the address field:

http://neo.mcs.utulsa.edu:16405/index.html

I will run a test on your server by creating an index.html file with references to gif, jpeg and pdf files to make sure that it can handle the required media types.

Note that your server is expected to run in any platform where the Java is supported (there might be issues with file path specification).

6 Report

You must include a well typed report explaining your design and implementation. You MUST implement the required features as described in this document and you MAY include any other additional features as long as they are HTTP/1.1 compliant.

7 Submission

The report and electronic submissions are due at class time on the due date via Harvey.

DUE DATE: December 1st, 2017 at 11:00am.