Introduction to Web Science

Assignment 5

Prof. Dr. Steffen Staab
staab@uni-koblenz.de

René Pickhardt
rpickhardt@uni-koblenz.de

Korok Sengupta
koroksengupta@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: November 30, 2016, 10:00 a.m.
Tutorial on: December 2, 2016, 12:00 p.m.

Please look at the lessons 1) **Dynamic Web Content** & 2) **How big is the Web?**

For all the assignment questions that require you to write code, make sure to include the code in the answer sheet, along with a separate python file. Where screen shots are required, please add them in the answers directly and not as separate files.

Team Name: XXXX

1 Creative use of the Hyptertext Transfer Protocol (10 Points)

HTTP is a request response protocol. In that spirit a client opens a TCP socket to the server, makes a request, and the server replies with a response. The server will just listen on its open socket but cannot initiate a conversation with the client on its own.

However you might have seen some interactive websites which notify you as soon as something happens on the server. An example would be Twitter. Without the need for you to refresh the page (and thus triggering a new HTTP request) they let you know that there are new tweets available for you. In this exercise we want you to make sense of that behaviour and try to reproduce it by creative use of the HTTP protocol.

Have a look at `server.py`[1] and `webclient.html` (which we provide). Extend both files in a way that after `webclient.html` is servered to the user the person controlling the server has the chance to make some input at its commandline. This input should then be send to the client and displayed automatically in the browser without requiring a reload. For that the user should not have to interact with the webpage any further.

1.1 webclient.html

```
1: <html>
2: <head>
3:         <title>Abusing the HTTP protocol - Example</title>
4: </head>
5: <body>
6:         <h1>Display data from the Server</h1>
7:         The following line changes on the servers command line
8:         input: <br>
9:         <span id="response" style="color:red">
10:                 This will be replaced by messages from the server
11:        </span>
12: </body>
13: </html>
```

1.2 Hints:

- This exercise is more like a riddle. Try to focuse on how TCP sockets and HTTP work and how you could make use of that to achieve the expected behaviour. Once you have an idea the programming should be straight-forward.
- The Javascript code that you need for this exercise was almost completely shown in one of the videos and is available on Wikiversity.

[1] you could store the code from http://blog.wachowicz.eu/?p=256 in a file called `server.py`

- In that sense we only ask for a "proof of concept" nothing that would be stable out in the wilde.
  - In particular, don't worry about making the server uses multithreading. It is ok to be blocking for the sake of this exercise.
- Without use of any additional libraries or AJAX framework we have been able to solve this with 19 lines of Javascript and 11 lines of Python code (we provide this information just as a way for you to estimate the complexity of the problem, don't worry about how many lines your solution uses).

Answer:

```
<html>
<head>
        <title>Abusing the HTTP protocol - Example</title>
</head>
<body onload="validator()">
<script>
        function validator(){
                        if (window.XMLHttpRequest) { // code for IE7+, Firefox, C
                                var ajax = new XMLHttpRequest();
                        }
                        else { // code for IE6, IE5
                                var ajax = new ActiveXObject("Microsoft.XMLHTTP")
                        }
                        ajax.open('GET', 'wait_for_server', true);
                        ajax.onreadystatechange = function(){
                                if (ajax.readyState == 4 && ajax.status == 200){
                                        document.getElementById("response").innerH
                                }
                        };
                        ajax.send();
        }
</script>
        <h1>Display data from the Server</h1>
        The following line changes on the servers command line
        input: <br>
        <span id="response" style="color:red">
                This will be replaced by messages from the server
        </span>
</body>
</html>
```

```
import socket   # Networking support
import signal   # Signal support (server shutdown on signal receive)
import time     # Current time
import os
```

```
 6: class Server:
 7:
 8:  def __init__(self, port = 80):
 9:      self.host = 'localhost'
10:      self.port = port
11:      self.www_dir = os.getcwd()
12:
13:
14:  def activate_server(self):
15:      self.socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
16:      try:
17:          self.socket.bind((self.host, self.port))
18:      except Exception as e:
19:          user_port = self.port
20:          self.port = 8080
21:
22:          try:
23:              self.socket.bind((self.host, self.port))
24:
25:          except Exception as e:
26:              self.shutdown()
27:              import sys
28:              sys.exit(1)
29:
30:      print ("Server successfully acquired the socket with port:", self.port)
31:      self._wait_for_connections()
32:
33:  def shutdown(self):
34:      try:
35:          s.socket.shutdown(socket.SHUT_RDWR)
36:      except Exception as e:
37:          print("Warning: could not shut down the socket. Maybe it was already clos
38:
39:
40:  def _gen_headers(self,  code):
41:
42:      header = ''
43:      if (code == 200):
44:          header = 'HTTP/1.1 200 OK\n'
45:      elif(code == 404):
46:          header = 'HTTP/1.1 404 Not Found\n'
47:      current_date = time.strftime("%a, %d %b %Y %H:%M:%S", time.localtime())
48:      header += 'Date: ' + current_date +'\n' + 'Server: Simple-Python-HTTP-Server'
    # signal that the conection wil be closed after complting the request
49:
50:      return h
51:
52:  def _wait_for_connections(self):
53:
```

```
54:     while True:
55:         print ("Awaiting New connection")
56:         self.socket.listen(3) # maximum number of queued connections
57:         conn, addr = self.socket.accept()
58:         data = conn.recv(1024) #receive data from client
59:         string = bytes.decode(data) #decode it to string
60:
61:         request_method = string.split(' ')[0]
62:         print ("Method: ", request_method)
63:         print ("Request body: ", string)
64:
65:         if (request_method == 'GET') | (request_method == 'HEAD'):
66:             file_requested = string.split(' ')
67:             file_requested = file_requested[1] # get 2nd element
68:             file_requested = file_requested.split('?')[0]  # disregard anything
69:
70:             if (file_requested == '/'):  # in case no file is specified by the b
71:                 file_requested = '/index.html' # load index.html by default
72:
73:             elif(file_requested != '/wait_for_server'):
74:
75:                 file_requested = self.www_dir + file_requested
76:                 print ("Serving web page [",file_requested,"]")
77:                 try:
78:                     file_handler = open(file_requested,'rb')
79:                     if (request_method == 'GET'):  #only read the file when GET
80:                         response_content = file_handler.read() # read file conte
81:                     file_handler.close()
82:
83:                     response_headers = self._gen_headers( 200)
84:
85:                 except Exception as e: #in case file was not found, generate 404
86:                     print ("Warning, file not found. Serving response code 404\n"
87:                     response_headers = self._gen_headers( 404)
88:
89:                     if (request_method == 'GET'):
90:                         response_content = b"<html><body><p>Error 404: File not
91:
92:                 server_response =  response_headers.encode() # return headers fo
93:                 if (request_method == 'GET'):
94:                     server_response +=  response_content  # return additional co
95:
96:
97:                 conn.send(server_response)
98:                 print ("Closing connection with client")
99:                 conn.close()
100:            else:
101:                inpt = input("server input: ")
102:                conn.send(inpt.encode())
```

```
103:             else:
104:                 print("Unknown HTTP request method:", request_method)
105:
106:
107: print ("Starting web server")
108: s = Server(80)  # construct server object
109: s.activate_server() # aquire the socket
```

2 Web Crawler (10 Points)

Your task in this exercise is to "crawl" the `Simple English Wikipedia`. In order to execute this task, we provide you with a mirror of the Simple English Wikipedia at `141.26.208.82`.

You can start crawling from `http://141.26.208.82/articles/g/e/r/Germany.html` and you can use the `urllib` or `doGetRequest` function from the last week's assignment.

Given below is the strategy that you might adopt to complete this assignment:

1. Download `http://141.26.208.82/articles/g/e/r/Germany.html` and store the page on your file system.
2. Open the file in python and extract the local links. (Links within the same domain.)
3. Follow all the links and repeat steps 1 & 2.
4. Repeat step 3 until you have no new pages to detect.

2.1 Hints:

- Before you start this exercise, please have a look at Exercise 3.
- Make really sure your crawler doesn't follow external urls to domains other than `http://141.26.208.82`. In that case you would start crawling the entire web
- Expect the crawler to run about XXX Minutes if you start it from the university network. From home your runtime will most certainly be even longer.
- It might be useful for you to have some output on the crawlers commandline depicting which URL is currently being fetched and how many URLs have been fetched so far and how many are currently on the queue.
- You can (but don't have to) make use of breadth-first search.
- It probably makes sense to take over the full paths from the pages of the Simple English Wikipedia and use the same folder structure when you save the html documents.
- You can (but you don't have to) speed up the crawler significantly if you use multithreading. However you should not use more than 10 threads in order for our mirror of Simple English Wikipedia to stay alive.

Answer:

```
1: import wget
2: import re
3: from urllib.parse import urlparse
```

```
 4: import urllib.request
 5: import socket
 6: import time
 7: import queue
 8:
 9: def doHTTPRequest(url):
10:
11:     o = urlparse(url)
12:     host = o.netloc
13:     path = o.path
14:     file_name = url.split('/')[-1]
15:     print(file_name)
16:     port = 80  # web
17:     urlQueue.put(url)
18:
19:     while not urlQueue.empty():
20:         print('\n # Creating socket')
21:         s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
22:
23:         print('# Getting remote IP address')
24:         remote_ip = socket.gethostbyname(host)
25:
26:         print('# Connecting to server, ' + host + ' (' + remote_ip + ')')
27:         s.connect((remote_ip, port))
28:
29:         print('# Sending data to server')
30:         request = "GET " + path + " HTTP/1.0\r\n\r\n"
31:         s.sendall(request.encode('utf'))
32:
33:         print('# Receive data from server \n ')
34:         HTMLcontent = recv_timeout(s)
35:         links = findURLS(HTMLcontent)
36:         downloadFILE(links, HTMLcontent)
37:
38: def recv_timeout(the_socket, timeout=2):
39:     # make socket non blocking
40:     the_socket.setblocking(0)
41:     # total data partwise in an string
42:     totalData=""
43:     data = '';
44:     # beginning time
45:     begin = time.time()
46:     while 1:
47:         # if we got some data, then break after timeout
48:         if totalData and time.time() - begin > timeout:
49:             break
50:
51:         # if you got no data at all, wait a little longer, twice the timeout
52:         elif time.time() - begin > timeout * 2:
```

```
53:             break
54:
55:         # recv something
56:         try:
57:
58:             data = the_socket.recv(8192)
59:             if data:
60:                 totalData = totalData + data.decode()
61:                 # change the beginning time for measurement
62:                 begin = time.time()
63:             else:
64:                 # sleep for sometime to indicate a gap
65:                 time.sleep(0.1)
66:         except:
67:             pass
68:     if(totalData.find("200 OK") != -1):
69:         splitted = str(totalData).split('\r\n\r\n')
70:         htmlbody = splitted[1]
71:     # return 0 for successfully executed function
72:     return htmlbody
73:
74:
75: def downloadFILE (links, HTMLcontent):
76:     i = 0
77:     while (i < len(links)):
78:         o = urlparse(links[i])
79:         hst = o.netloc
80:         if (hst == "141.26.208.82"):
81:             fnn = links[i].split('/')[-1]
82:             file_ = open(fnn, 'w')
83:             file_.write(str(HTMLcontent.encode('utf-8')))
84:             file_.close()
85:             urlQueue.put(links[i])
86:         i = i + 1
87:
88: def findURLS(content):
89:     RegExpr = '<a href="?\'?([^"\'>]*)'
90:     links = re.findall(RegExpr, content)
91:     print(len(links))
92:     i = 0  # printing and download the images
93:     j=0
94:     while (i < len(links)):
95:         links[i] = urllib.parse.urljoin("http://141.26.208.82", str(links[i]))
96:         i = i + 1
97:
98:     return links
99:
100:
101: f = 'http://141.26.208.82/articles/g/e/r/Germany.html'
```

```
102: urlQueue = queue.Queue()
103: doHTTPRequest(f)
```

3 Web Crawl Statistics (10 Points)

If you have successfully completed the first exercise of this assignment, then please provide the following details. You may have to tweak your code in the above exercise for some of the results.

3.1 Phase I

1. Total Number of *webpages* you found.
2. Total number of links that you encountered in the complete process of crawling.
3. Average and median number of links per web page.
4. Create a *histogram* showing the distribution of links on the crawled web pages. You can use a bin size of 5 and scale the axis from 0-150.

3.2 Phase II

1. For every page that you have downloaded, count the number of internal links and external links.
2. Provide a *scatter plot* with number of internal links on the X axis and number of external links on the Y axis.

Answer:

```
1: import socket
2: import os
3: def url_processor(Input):
4:     protocol, port, path, Frag, param= []
5:     index, Flag = 0
6:     if "://" in Input:
7:         Flag = 1
8:     while index < len(Input) and Flag == 1 and Input[index] != ':':
9:         protocol.append(Input[index])
10:        index += 1
11:    if Flag == 1:
12:        index += 1
13:    while Input[index] == '/' or Input[index] == ' ':
14:        index += 1
15:    j = index
16:    while index < len(Input) and Input[index] not in [':', '/']:
17:        index += 1
18:    portflag = 0
19:    if index < len(Input) and Input[index] == '/':
```

```
20:         portflag = 1
21:     index += 1
22:     names = Input[j:index - 1]
23:     names = names.split(".")
24:     while index < len(Input) and Input[index] != '/' and portflag == 0:
25:         port.append(Input[index])
26:         index += 1
27:     if portflag == 0:
28:         index += 1
29:     while index < len(Input) and Input[index] != '?':
30:         path.append(Input[index])
31:         index += 1
32:     index += 1
33:     while index < len(Input) and Input[index] != '#':
34:         param.append(Input[index])
35:         index += 1
36:     index += 1
37:     while index < len(Input):
38:         Frag.append(Input[index])
39:         index += 1
40:     return protocol, names, port, path, param, Frag
41:
42: def Header_seperator(file_name, string):
43:     try:
44:         content = open(file_name, "wb")
45:         splited_str = string.split(b"\r\n\r\n", 1)
46:         header = splited_str[0]
47:         header = header.split(b' ')
48:         if header[1] != b'200':
49:             return -1
50:         splited_str[1] = splited_str[1].replace(b'\\n', b'')
51:         splited_str[1] = splited_str[1].replace(b'\\t', b'')
52:         content.write(splited_str[1])
53:         content.close()
54:         return 0
55:     except Exception as ex:
56:         print("header_seperator : ", ex)
57: def response(socket, file_name):
58:     answer = bytearray()
59:     try:
60:         tmp = socket.recv(8196)
61:         while tmp != b'':
62:             for char in tmp:
63:                 answer.append(char)
64:             tmp = socket.recv(8196)
65:         socket.close()
66:         flag = Header_seperator(file_name, answer)
67:         return flag
68:         print("Task Finished")
```

```
69:     except Exception as ex:
70:         print("response ", ex)
71: def file_to_string(file_name):
72:     file = open(file_name, "r")
73:     string = ""
74:     for item in file:
75:         string += item
76:     file.close()
77:     return string
78: def string_parser(string, start, end):
79:     index = string.find(start)
80:     answer = []
81:     while True:
82:         if index == -1:
83:             break
84:         end_index = string[index + len(start):].find(end)
85:         i = index + len(start) + end_index + len(end)
86:         answer.append(string[index:i])
87:         string = string[index + len(start) + end_index + len(end):]
88:         index = string.find(start)
89:     return answer
90: def client(socket, url, file_name):
91:     protocol, names, port, path, param, Frag = url_processor(url)
92:     protocol = "".join(protocol)
93:     if len(protocol) != 0 and protocol.upper() != "HTTP":
94:         print("This is HTTP Client program!")
95:         return
96:     server = ""
97:     for name in names:
98:         server += name + "."
99:     path = "".join(path)
100:    server = server[0:len(server) - 1]
101:    try:
102:        socket.connect((server, 80))
103:    except:
104:        print("not connected ")
105:    if len(path) == 0:
106:        get = b"GET /HTTP/1.1\r\n\r\n "
107:        print(get)
108:    else:
109:        get = b"GET /" + path.encode() + b" /HTTP/1.1\r\n\r\n"
110:    try:
111:        socket.sendall(get)
112:    except:
113:        print("data can not be sent!")
114:    flag = response(socket, file_name)
115:    return flag
116: def extract(file_name):
117:    try:
```

```
118:         links = []
119:         tmp = ""
120:         str_temp = file_to_string(file_name)
121:         i = 1
122:         index = 0
123:         split_data = str_temp.split("<a")
124:         while i < len(split_data):
125:             string = split_data[i]
126:             index = string.find('href="')
127:             if index != -1:
128:                 index += len('href="')
129:                 while string[index] != '"':
130:                     tmp += string[index]
131:                     index += 1
132:                 links.append(tmp)
133:                 tmp = ""
134:             i += 1
135:         return links
136:     except:
137:         return []
138: def merg_3(src, dest, array):
139:     for item in src:
140:         if item not in array.keys():
141:             dest.append(item)
142:             array[item] = 1
143: def correct_url(string):
144:     i = 0
145:     while i < len(string):
146:         if string[i] in ['.', '/']:
147:             i += 1
148:         else:
149:             break
150:     return string[i - 1:]
151: def get_request(url, file_name):
152:     import socket
153:     socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
154:     flag = client(socket, url, file_name)
155:     return flag
156: socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
157: url = "http://141.26.208.82/articles/g/e/r/Germany.html"
158: get_request(url, "pages/level_0_1.html")
159: all_links = {}
160: Total_number_of_pages = 0
161: level = 0
162: Internal_pages_count = {}
163: external_pages_count = {}
164: while True:
165:     n = 1
166:     file_name = "pages/level_" + str(level) + "_" + str(n) + ".html"
```

```
167:     current_links = []
168:     while os.path.isfile(file_name):
169:         temp = extract(file_name)
170:         merg_3(temp, current_links, all_links)
171:         n += 1
172:         Total_number_of_pages += 1
173:         file_name = "pages/level_" + str(level) + "_" + str(n) + ".html"
174:     index = 1
175:     print(len(current_links))
176:     if len(current_links) == 0:
177:         break
178:     for link in current_links:
179:         if link[0] in ['.', '/']:
180:             url = correct_url(link)
181:             url = "http://141.26.208.82" + url
182:             file_name = "pages/level_" + str(level + 1) + "_" + str(index) + ".htm
183:             flag = 0
184:             flag = get_request(url, file_name)
185:             if flag == 0:
186:                 index += 1
187:     level = level+ 1
```

Important Notes

Submission

- Solutions have to be checked into the github repository. Use the directory name `groupname/assignment5/` in your group's repository.
- The name of the group and the names of all participating students must be listed on each submission.
- Solution format: all solutions as *one* PDF document. Programming code has to be submitted as Python code to the github repository. Upload *all* `.py` files of your program! Use `UTF-8` as the file encoding. *Other encodings will not be taken into account!*
- Check that your code compiles without errors.
- Make sure your code is formatted to be easy to read.
  - Make sure you code has consistent indentation.
  - Make sure you comment and document your code adequately in English.
  - Choose consistent and intuitive names for your identifiers.
- Do *not* use any accents, spaces or special characters in your filenames.

Acknowledgment

This latex template was created by Lukas Schmelzeisen for the tutorials of "Web Information Retrieval".

LaTeX

Currently the code can only be build using LuaLaTeX, so make sure you have that installed. If on Overleaf, there's an error, go to settings and change the LaTeXengine to `LuaLaTeX`.